THE EMPIRE ON EVERY PAGE

AN EXPLORATION OF IMPERIALISM IN THE VICTORIAN PRESS USING DIGITAL METHODOLOGIES, 1850-1900

Submitted in partial fulfilment of the degree of Doctor of Philosophy (PhD)

16/7/2020

Quintus van Galen
Edge Hill University

Abstract

Quintus van Galen, Submitted in partial fulfilment of the degree of Doctor of Philosophy (PhD), The Empire On Every Page: An Exploration of Imperialism in the Victorian Press using Digital Methodologies, 1850-1900.

As historical research increasingly involves digitised sources and methods, historians have to adapt their practice to this 'digital turn'. This thesis explores the implications of this reality by building two tools to use on the *British Library Nineteenth-Century Newspapers* archive, in order to establish the viability of researcher-developed tools. With this it will carry out its investigation in the prevalence of banal imperialism in the nineteenth century. The first tool will use Topic Modelling, the second will visualise the spatial nature of newspaper articles. For both these the thesis will establish the viability and limitations in service of the historical method. It shows that while topic models are useful for general exploration, on this archive they lack the required accuracy to anchor arguments. It finds the visualisation tool it developed allows for the discovery of spaces within a newspaper dedicated to given topics. This tool also allows the study of the development of layout and design practices on historical newspapers in new ways. It concludes that both tools are best used as aides for a human researcher, but cautions against allowing any tool to supplant human interpretation. Its findings on imperialism support the available literature. It establishes a wide body of banal imperial references permeated everyday Victorian life, and especially the press. It also establishes some aspects of theoretical frameworks for nationalism are applicable to studying imperialism.

Dedication

Many people helped make this thesis possible, but as always there are some who it would not have been possible to complete without. First, my supervision team, Dr. Alyson Brown, Dr. Mark Hall, and Dr. Bob Nicholson, who supported me throughout the project and guided the multitude of wild ideas into something resembling structure. I also wish to thank my internal and external examiners for the registration and progression viva, Dr. Sarah Irving and Dr. Melodee Beals, who comments contributed to the direction of the thesis. I would also like to thank Dr. James Mussell and Dr. Paul Ward for agreeing to give their time to act as members of the examination panel.

On a more personal level, this thesis was a journey of discovery in more ways than one, for which I took a chance and moved to a country about to leave a union of nations. I owe my sanity at the end of this to those that travelled with me, the other PhD candidates in the department, and in specific Dr. Gemma Outen, Dr. Phillipa Holloway, Dr. Padric Soulsby and Dr. Phil Rawsthorne. They provided much-needed relief in times that everything seemed to go wrong and nothing seemed to work, and someone to moan to about computers that never did what I wanted them to. I also owe a debt of gratitude to Mr. Gerard Maters, who provided me with powerful computing hardware when I needed it most. Finally, I would like to thank my parents, who put up with me claiming tablespace with folders and drafts. They have finally been rewarded in their belief that their son might someday be a writer.

Table of Contents

Table of Figures

Introduction

Now it must be seen that the stone thus brought under the artist's hand to the beauty of form is beautiful not as stone- for so the crude block would be as pleasant- but in virtue of the form or idea introduced by the art. This form is not in the material; it is in the designer before ever it enters the stone; and the artificer holds it not by his equipment of eyes and hands but by his participation in his art.

– Plotinius, *Ennead*, V.8.1

Though a gulf of eighteen centuries separate the writing of these words by Plotinius and this thesis, it is still fitting to open with them. For his words are equally applicable to the work of the sculptor transforming marble, as they are to the historian working with data. While 'raw' data may contain meanings in and of itself, and be the product of many interpretations and processes – what Couper has dubbed 'paradata' – it only gains a historical meaning after the attentions of the historian are lavished on it with the intent to interpret it.[1] This thesis comes to a similar conclusion in its search for ways to apply digital techniques to historical research. While it can be a magnificent chisel, that is still all it can ever be – a tool to help the historian ply their trade.

This is pertinent because we live in a time when data, not sources, are becoming increasingly prominent in historical research.[2] Instead of going to the

[1] Mick Couper, Frauke Kreuter, and Lars Lyberg, 'The Use of Paradata to Monitor and Manage Survey Data Collection', in *JSM Proceedings - Survey Research Methods Section* (presented at the American Statistical Association, Alexandria, VA: American Statistical Association, 2010), pp. 281–96 <http://www.asasrms.org/Proceedings/y2010/Files/306107_55863.pdf>.

[2] Joris van Eijnatten, Toine Pieters, and Jaap Verheul, 'Big Data for Global History: The Transformative Promise of Digital Humanities', *BMGN - Low Countries Historical Review*, 128.4 (2013), 55–77 <https://doi.org/10.18352/bmgn-lchr.9350>; Gertjan Willems, 'Digitale Tools Voor Kwalitatieve Data-

British Library's reading rooms, many historians now access nineteenth-century newspapers in a digitised form where possible.[3] It is certainly much more convenient: no more long train journeys into London, no more relying on large and unwieldy indices available for just one or two papers, and instead of waiting for a librarian to bring the paper you've requested to the reading room, it's delivered right to your office in the blink of an eye. The nineteenth-century newspaper as quick and easy as a modern news website.[4] Paradoxically, this ease of access provides a challenge to historians: a profession that for centuries has devised methods to deal with a scarcity of information, suddenly has to handle information overload. The question then becomes: can the modern historian, even when working on old sources, still afford to rely only on manual methods?

This thesis will argue this is not the case. Some historians have explored the new reality in depth, such as Bob Nicholson, James Mussell, Paul Fyfe and others.[5] Yet overall, there appears to be too little critical engagement with the enormous implications of the rise of the digital archive – and associated digital scholarship – has on historical methods. This thesis will build upon existing work by exploring

---

Analyse Binnen Historisch Comminicatiewetenschappelijk Onderzoek: Toepassingen En Reflecties', *Tijdschrift Voor Communicatiewetenschap*, 45.3 (2017), 170–83.

[3] Patrick Leary, 'Googling the Victorians', *Journal of Victorian Culture*, 10.1 (2005), 72–86 <https://doi.org/10.3366/jvc.2005.10.1.72>.

[4] Janine Solberg, 'Googling the Archive: Digital Tools and the Practice of History', *Advances in the History of Rhetoric*, 15.1 (2012), 53–76 <https://doi.org/10.1080/15362426.2012.657052>.

[5] Bob Nicholson, 'The Digital Turn', *Media History*, 19.1 (2013), 59–73 <https://doi.org/10.1080/13688804.2012.752963>; James Mussell, *The Nineteenth-Century Press in the Digital Age*, Palgrave Studies in the History of the Media (Basingstoke: Palgrave Macmillan, 2012); Paul Fyfe, 'Access, Computational Analysis, and Fair Use in the Digitized Nineteenth- Century Press', *Victorian Periodicals Review*, 51.4 (2018), 716–37; Hinke Piersma and Kees Ribbens, 'Digital Historical Research: Context Concepts and the Need for Reflection', *BMGN-Low Countries Historical Review*, 124.4 (2013), 78–102; Jeurgens, Charles, 'The Scent of the Digital Archive: Dilemmas with Archive Digitisation', *BMGN-Low Countries Historical Review*, 128, 2013, 30–54; Adrian Bingham, '"The Digitization of Newspaper Archives: Opportunities and Challenges for Historians"', *Twentieth Century British History*, 21.2 (2010), 225–31 <https://doi.org/10.1093/tcbh/hwq007>; Solberg.

both the general problem of integrating historical and traditional methods, and answering these questions of integration for two tools: topic modelling, a popular technique using Latent Dirichlet Allocation; and spatial newspaper visualisation, which this project developed itself.

Methodology: Integration

First, this thesis asks how the use of digital tools and digitised sources could be integrated in 'traditional' historical methodology, and what models can be used to understand this blending of methodological traditions. This question is intrinsically tied to the development of the field of the Digital Humanities (DH). Since the turn of the millennium, DH has taken flight as the interdisciplinary umbrella under which computational methods are used to undertake humanities research. The discipline itself argues that it is its own field of study, with its own methodological and epistemological practices. This thesis argues that while such an approach may be suitable for other humanities fields, it is not conducive to the production of historical research. Instead, it proposes that in the context of the production of historical knowledge, the Digital Humanities are best understood as an ancillary science to history. DH provides the methods by which the research takes place, but the setting of the questions and the epistemological framework within which the answers are to be understood is historical. Academic history also possesses the models for working with ancillary sciences, while still being respectful to these fields as their own domains of knowledge and practice.

This argument prompts three subsequent questions directly related to the integration of the digital and the humanities. The first of these seeks to establish

what the role of source criticism should be when a historian uses digital archives and digital tools. Of course, this project is not the first to posit this question. James Mussell has pointed out that a historian working on a digitised archive should familiarise themselves with the history of their chosen archive, to the point that they can understand how the archive decides what material they are presented with, and where the material that resides in the archive has originated.[6] The end goal of this process of archival criticism is understanding the way the history of the archive influences and limits the data it contains. This also places restrictions on what can be done with and concluded from the archive's contents. This thesis agrees with Mussell's emphasis on source criticism, and discusses the importance of critiquing not only the digital archive, but also the physical archive from which it was generated. It discusses how events in the history of the archive, the *British Library Nineteenth-Century Newspapers I* and *II* datasets, impact the form of the digital sources.

The second question prompted by an attempt to integrate historical and computational practices is related to computational limitations. Other scholars, such as Stephen Ramsay, have discussed the general ways in which humanities research can be limited by access to computational resources.[7] This thesis shows that the constraints imposed by these resources have significant impact on the design of a research project. By throttling the speed at which data retrieval operations take place, computational constraints limit the possibility to use

[6] James Mussell, *The Nineteenth-Century Press in the Digital Age*, pp. 195–97.

[7] Stephen Ramsay, 'High Performance Computing for English Majors', 2008 <http://stephenramsay.us/text/2008/04/14/high-performance-computing-for-english-majors/> [accessed 21 October 2016].

keywords that may not result in an adequate number of relevant sources being discovered. The limits they impose on the number of sources that can be analysed mean there are limits to the comparative capabilities of a research project. Both of these limiting factors force the researcher to be conservative in their design.

Having established that source criticism is a crucial aspect of integrating digital methods with historical practices, and that computational resources impact the scale and scope of such blended research projects, the third question this thesis explores concerns criticism of the tools themselves. Work by Koolen, van Gorp and van Osenbrugge on the theoretical foundation of tool criticism stresses the importance of understanding the process by which a tool transforms data, but offers few handles on how to gain such an understanding.[8] This thesis follows the argument by Rockwell and Ramsay, that building a tool produces a scholarly object, which can intertwine with critical discourse.[9] Thus, this research project establishes the plausibility of building tools oneself, as a single researcher, and shows that doing so offers significant benefits. It shows that by building tools the researcher gains a deep understanding of the tool's operation, as well as the ability to match the tool's intake of data to the particularities of the archive. Additionally, building tools also allows the results of the transformations of data to be presented in a way that respects historical research practices.

[8] Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen, 'Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice', *Digital Scholarship in the Humanities*, 34.2 (2019), 368–85 (pp. 8; 21) <https://doi.org/10.1093/llc/fqy048>.

[9] Stephen Ramsay and Geoffrey Rockwell, 'Developing Things: Notes towards and Epistemology of Building in the Digtital Humanities', in *Debates in the Digital Humanities*, ed. by Matthew K. Gold (Minneapolis: University of Minnesota Press, 2012) <https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities>.

Methodology: Topic Modelling

Having made these contributions to the debates on integrating traditional and digital methods in the context of historical research, this thesis will consider two specific research methods, and develop ways to sensibly use them. It uses these two case studies as means to test its broader methodological approach, but will also address specific issues related to them. The first of these tools is topic modelling, specifically in the form of Latent Dirichlet Allocation (LDA). This has recently been gaining popularity amongst scholars as an application for textual classification: in 2000 there were 22 articles about topic models indexed on Google Scholar; by 2018 this number had risen to 3,270.[10] Gale-Cengage, the purveyor of the *Times Digital Archive, 1785-2004* and other, similar resources, integrated topic modelling into their Digital Scholar Lab in 2018.[11] This greatly expands the number of historians that have access to these tools. Thus, with the rising popularity of topic modelling as a method, there is a need for research into the structural application of topic models, and their methodological and epistemological implications. After all, if a tool is just a button-click away it becomes tempting to use, even if the processes behind it may be poorly understood.[12]

Topic models have issues with being under-theorised, and have not yet been successfully made to work in a structurally, methodologically and epistemologically

---

[10] John W. Mohr and Petko Bogdanov, 'Introduction—Topic Models: What They Are and Why They Matter', *Poetics*, 41.6 (2013), 545–69 <https://doi.org/10.1016/j.poetic.2013.10.001>; René Brauer, Mirek Dymitrow, and Mats Fridlund, 'The Digital Shaping of Humanities Research: The Emergence of Topic Modeling within Historical Studies', *Enacting Futures: DASTS 2014*, 2014.

[11] Gale-Cengage, 'New Product Enhancements Announced for Digital Scholar Lab', *Product Support*, 2019 </updates/dsl-feb19> [accessed 3 March 2020].

[12] For a general discussion of this issue, see: Ted Underwood, 'Theorizing Research Practices We Forgot to Theorize Twenty Years Ago', *Representations*, 127.1 (2014), 64–72 <https://doi.org/10.1525/rep.2014.127.1.64>.

sound way. The number of times they have been employed in a historical context is legion, however, they have only rarely been used in a way that saw them embedded at the heart of a historical method.[13] Often times the tool (topic modelling), took centre stage over the research object, leading to projects that were more interested in advancing the technology of topic modelling than answering a historical question.[14] This has been the case since the very first presentation of the topic modelling algorithm LDA, which was tested on the *New York Times* corpus.[15] This project contributes to the theoretical foundation of LDA by defining the limits within which it can be used on the *British LibraryNewspapers, 1800-1900* .

To establish these limits, this thesis first establishes how topic modelling relates itself to historical epistemology. There has been little work done to fully integrate topic modelling into historiographical, or even humanistic, practice despite multiple calls to do so.[16] Mostly, this has taken place within the spheres of

[13] See for two notable exceptions: David J. Newman and Sharon Block, 'Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper', *Journal of the American Society for Information Science and Technology*, 57.6 (2006), 753–67; Maarten van den Bos and Hermione Giffard, 'Mining Public Discourse for Emerging Dutch Nationalism', 10.3 (2016) <http://www.digitalhumanities.org/dhq/vol/10/3/000263/000263.html#vaneijnatten2014> [accessed 9 October 2017]. The latter explicitly mentions these issues in the discussion.

[14] Andrew J. Torget and others, 'Mapping Texts: Combining Text-Mining and Geo-Visualization to Unlock the Research Potential of Historical Newspapers', *University of North Texas Digital Library*, 2011 <https://pdfs.semanticscholar.org/4b40/d6b77b332214eefc7d1e79e15fbc2d86d86a.pdf> [accessed 26 September 2017]; Tze-I Yang, Andrew J. Torget, and Rada Mihalcea, 'Topic Modeling on Historical Newspapers', in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Stroudsburg, PA, USA: Association for Computational Linguistics, 2011), pp. 96–104 <http://dl.acm.org/citation.cfm?id=2107636.2107649> [accessed 26 September 2017]; David Hall, Daniel Jurafsky, and Christopher D. Manning, 'Studying the History of Ideas Using Topic Models', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2008), pp. 363–371 <http://dl.acm.org/citation.cfm?id=1613763> [accessed 26 September 2017]; Nina Tahmasebi, 'A Study on Word2Vec on a Historical Swedish Newspaper Corpus', in *DHN*, 2018; Paul DiMaggio, Manish Nag, and David Blei, 'Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding', *Poetics*, 41.6 (2013), 570–606 <https://doi.org/10.1016/j.poetic.2013.08.004>.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3.Jan (2003), 993–1022.

[16] René Brauer, Mirek Dymitrow, and Mats Fridlund, 'The Digital Shaping of Humanities Research: The Emergence of Topic Modeling within Historical Studies', in *Enacting Futures: DASTS 2014*, 2014; Maarten van den Bos and Hermione Giffard, 'Mining Public Discourse for Emerging Dutch Nationalism', 10.3

discourse studies, by scholars seeking to theorise the integration with linguistics, meaning a historical integration is needed more than ever. This thesis argues that there exists a problematic relationship between topic models and historical praxis. Topic models operate by reducing a text to a frequency count of its constituent words, in essence disassembling the source into textual components. By contrast, the paradigm of the historical method since the nineteenth century has been to seek as original a form of the source as possible when conducting historical research. Consequently, this thesis shows that the uses of topic models for historical research are limited to cases where the goal is to make general observations about the corpus.

However, the answers to this first question on the use of LDA give rise to a second question: what limits are imposed on the use of topic models of nineteenth-century newspapers by the sources itself, and how does this affect their usefulness? While several authors have established the connection between document OCR transcription errors and topic model quality, their findings offer little practical implications for historical newspaper research under such circumstances.[17] This thesis shows the impact transcription errors have on the

(2016) <http://www.digitalhumanities.org/dhq/vol/10/3/000263/000263.html#vaneijnatten2014> [accessed 9 October 2017]; Thomas Jacobs and Robin Tschötschel, 'Topic Models Meet Discourse Analysis: A Quantitative Tool for a Qualitative Approach', *International Journal of Social Research Methodology*, 22.5 (2019), 469–85 <https://doi.org/10.1080/13645579.2019.1576317>.

[17] Daniel D. Walker, William B. Lund, and Eric K. Ringger, 'Evaluating Models of Latent Document Semantics in the Presence of OCR Errors', in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10 (Cambridge, Massachusetts: Association for Computational Linguistics, 2010), pp. 240–250; Daniel Walker, Eric Ringger, and Kevin Seppi, 'Evaluating Supervised Topic Models in the Presence of OCR Errors', ed. by Richard Zanibbi and Bertrand Coüasnon (presented at the IS&T/SPIE Electronic Imaging, Burlingame, California, USA, 2013), p. 865812 <https://doi.org/10.1117/12.2008345>; Stephen Mutuvi and others, 'Evaluating the Impact of OCR Errors on Topic Modeling', in *Maturity and Innovation in Digital Libraries*, ed. by Milena Dobreva, Annika Hinze, and Maja Žumer, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2018), pp. 3–14 <https://doi.org/10.1007/978-3-030-04257-8_1>.

usefulness of the resulting models, limiting them to fewer topics than current literature suggests.

Using the answers to these two questions, this thesis defines the practical limits of topic modelling when used to answer historical questions. It established the kinds of research question that these tools are capable and incapable of answering. It finds topic models are useful tools for high-level exploration of a corpus or keyword-generated subset of a corpus, when a broad classification of the sources is all that is needed. This can practically be used to chart the rise and fall of journalistic genres, and to explore the contexts in which a keyword was used. They are also useful for posing questions based on the appearance, or non-appearance, of topics, which may contradict expected results based on established historiography.

Methodology: Spatial Visualisation

The second main contribution of this thesis focuses on an entirely new tool, designed to explore the spatial aspect of newspaper articles. While digital archives are useful in their own right, they do impose a significant loss of information on the sources within them, taking away their physicality. As a result, those aspects of the source that are tied to its physical nature, such as the texture of the paper or the pattern of repetition of articles in space of a bound volume of newspapers, are taken away from the researcher's options for interpretation. These patterns are extremely important to understand the identity of newspapers and other periodicals, as the way its editor decided it should be put together and where each

(repeating) piece of content should be placed shapes the title's self.[18] The loss of these physical patterns impedes the understanding of historical newspapers.

This project has developed a means to recreate some of the physicality lost in the digitisation process by visualising the spatial nature of newspaper articles. It provides a way to recreate the repetition of the placement of articles between multiple issues of newspapers. The method it developed relies on the positional data used to guide keyword searches in the articles. By collating the frequency with which coordinate-page pairs occur and mapping it on a two-dimensional density plot (heatmap) it can show where in a paper any given discussion takes place. It thus allows for the exploration of that which topic models do not: the non-textual context of the sources.

In this specific case, there is a theoretical foundation for this focus on spatiality, as newspaper articles are more than just a self-contained collection of words; their position in a newspaper, their juxtaposition with surrounding articles, and the repeating patterns of the periodical's layout, all shape their meaning.[19] This thesis shows that a tool to explore these patterns hidden within periodical archives can be built, relying on a heatmap to show the article density of certain keyword-searched articles over time. It relies on the creative re-use of data, as it uses the article position data intended to highlight search results to count the number of articles that occupy the same space.

---

[18] Kristof van Gansen, "'Une Page Est Une Image.": Tekst Als Beeld in Arts et Métiers Graphiques', *Tijdschrift Voor Tijdschriftstudies*, 35, 2014, 5–21.

[19] See for early work theorising this: M. Pastoureau, 'L'illustration Du Livre: Comprendre Ou Rêver?', in *Histoire de l'édition Française. Tome I. Le Livre Conquérant. Du Moyen Âge Au Milieu Du XVIIe Siècle*, ed. by R. Chartier and H.J. Martin (Paris: Fayard, 1989), pp. 602–28; J.G. Lapacherie, 'De La Grammatextualité', *Poetique*, 59 (1984), 283–94.

Having developed the tool on a technical level, this thesis then has to address its use on a historiographic level: how does it relate itself to historical epistemology? As there is no existing literature on the method or tool that have just been developed, this thesis has to establish the foundation for the method itself. It finds the tool as developed has no inherent epistemological issues, as it relies on the best information available. It does not transform data, but instead counts a part of it that would be difficult, but not impossible, for a human to measure. However, there are epistemological choices that have to be made when using the tool, such as the time range, title, keywords used for article selection, and respect to pagination that are relevant, but these need be justified by the user and do not reflect on the inherent epistemological basis of the method.

Having established the fundamental epistemological soundness of the spatial visualisation of newspaper articles, this thesis defines the limits of its use for studying historical expressions of imperialism. These are still very narrow and limited only to this specific research question, but offer an indication of where future researchers may deploy the method. This thesis identifies three key limitations in its spatial visualisation of article placement: (1) the availability and accuracy of the data in the archive, (2) the availability of computational resources, and (3) the requirement for a keyword that produces a reliable selection of articles.

Historiography

When attempting to integrate historical and digital methodologies the researcher has to walk a tightrope between computer science and history.[20] It is tempting to commit fully to a computational approach, especially when a large part of the project involves building tools and testing their limits. However, that would risk losing sight of the historical research questions that initiated this project. This question, of where humanities end and computation begins, is one that has led to vigorous debate in the Digital Humanities, and is still unresolved.[21] In order to keep the balance between these two, this project develops its tools with specific historical research questions in mind.

It also means that this thesis will make a historical contribution to knowledge in its own right. In its case studies, it will engage with the historiography of the British Empire, specifically the theories of Michael Billig and John MacKenzie's Manchester School; respectively, banal nationalism and cultural imperialism.[22] Billig sought to explain the spread and construction of Nationalism through the everyday, subconscious and banal interactions a citizen has with their state. This theory has found use by historians of empire to understand the level to

[20] Stefan Jänicke, 'Valuable Research for Visualization and Digital Humanities: A Balancing Act', in *1st Workshop on Visualization for the Digital Humanities* (presented at the IEEE VIS 2016, Baltimore, USA, 2016) <http://vis4dh.dbvis.de/papers/2016/Valuable%20Research%20for%20Visualization%20and%20Digital%20Humanities%20A%20Balancing%20Act.pdf>.

[21] Alan Liu, 'The Meaning of the Digital Humanities', *PMLA*, 128.2 (2013), 409–23 <https://doi.org/10.1632/pmla.2013.128.2.409>; David M. Berry, 'The Computational Turn: Thinking About the Digital Humanities', *Culture Machine*, 12 (2011), 23; Stephen Ramsay, 'Who's In and Who's Out', 2011 <http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/> [accessed 21 October 2016]; Tito Orlandi, 'Reflections on the Development of Digital Humanities' (presented at the DH2019 - Busa Lecture, Utrecht, 2019).

[22] Michael Billig, *Banal Nationalism* (London: SAGE, 1995); John M. MacKenzie, *Propaganda and Empire: The Manipulation of the British Public Opinion 1880-1960* (Manchester: Manchester University Press, 1984).

which the empire permeated society both in the colonies and in Britain.[23] MacKenzie had similar goals when trying to understand the way British society was permeated with imperial symbolism and art. The case studies in this thesis test their theories of imperial penetration in British newspapers. The thesis finds the tools it developed indicate a wide body of imperial 'flags' in the British press, supporting the argument of the Manchester school for the existence of popular and banal imperialism.

This enquiry limits itself to a single archive: the *British Library Nineteenth Century Newspapers part I and II*. This dataset was created in 2003 by the British Library through the digitisation of the available print runs of 69 newspapers titles held in their collection for the period 1800-1900. It contains approximately 3 million pages of automatically transcribed text.[24] This project will further limit this range to the period between 1850 and 1900 for two reasons. The first is technical: the transcription of the later newspapers is generally better than that of earlier material. This makes it easier to use the tools developed. The second reason is historical. By beginning in 1850, we include events such as the Crimean War and

[23] See for example: Jeremy Black, review of *Visions of Empire: Patriotism, Popular Culture and the City, 1870–1939; Britain's Imperial Muse: The Classics, Imperialism, and the Indian Empire, 1784–1914; Crisis in the Mediterranean: Naval Competition and Great Power Politics, 1904–1914; Mapping the End of Empire: American and British Strategic Visions in the Postwar World; The Second British Empire: In the Crucible of the Twentieth Century; Distant Strangers: How Britain Became Modern*, by Brad Beaven and others, *Journal of World History*, 26.2 (2016), 395–400 <https://doi.org/10.1353/jwh.2016.0041>; N. C. Fleming, review of *Britain's Experience of Empire in the Twentieth Century*, by Andrew Thomson, *The Journal of Imperial and Commonwealth History*, 41.3 (2013), 536–37 <https://doi.org/10.1080/03086534.2013.823745>; Beth Mudford, 'Royal Celebrations in the Twenty-First Century: "Cool Britannia" versus "Britannia Ruled the Waves"', in *Identity Discourses and Communities in International Events, Festivals and Spectacles*, ed. by Udo Merkel, Leisure Studies in a Global Era (London: Palgrave Macmillan UK, 2015), pp. 116–34 <https://doi.org/10.1057/9781137394934_6>; James D Sidaway, 'The Dissemination of Banal Geopolitics: Webs of Extremism and Insecurity', *Antipode*, 40.1 (2008), 2–8 <https://doi.org/10.1111/j.1467-8330.2008.00568.x>.

[24] *British Library Newspapers* (Gale Cengage Learning) <https://www.gale.com/binaries/content/assets/gale-us-en/primary-sources/newsvault/gps_newsvault_britishlibrarynewspapers_academic_brochure_all.pdf>.

the Indian Mutiny, as well as the repeal of the taxes on newspapers. While we miss out on earlier events such as the coronation of Queen Victoria, the 1850 date offers a good balance between transcription quality and historical relevance.

This periodisation gives rise to a second historiographical contribution. This thesis investigates the applicability of the tools it integrates — topic modelling and spatial visualisation — to verify the applicability of theoretical frameworks outside of the parameters they were originally intended for. In this case, it will look specifically at Paul Ward's framework for understanding British national identity from 1870 onwards. He identifies five facets of Britishness: (1) the Urban-Rural divide that forced people to define their local allegiances; (2) Gender and the renegotiation of women's roles in public life; (3) the royal family and the empire; (4) Political allegiances; and (5) leisure, entertainment, and sports. [25] In its case studies, this thesis shows that some of these aspects may be applied to the study of Imperialism from 1850 onwards, while for others there is inconclusive evidence when using topic modelling and spatial visualisation.

Taken together, the methodological and historiographical contributions construct a greater narrative of the role tools play in our research, and the ways that we as researchers interact with our tools. This thesis emerged as a reaction against the 'dataism' of the Digital Humanities: a sense that if we had more data, faster computers, and better algorithms, we could come to a more complete insight about the human condition.[26] It takes as a given that there will always be questions

[25] Paul Ward, *Britishness Since 1870* (London and New York: Routledge, 2004), pp. 1–14.
[26] The term dataism is from: Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (New York: HarperCollins Publishers Ltd, 2016).

tools cannot answer; limits imposed on them by the data they ingest and the parameters within which their algorithms are designed to function.

Thesis Structure

Based on the four main questions of (1) methodological integration, the uses of respectively (2) topic modelling and (3) spatial visualisation, and (4) historiographic contributions to the debates on public imperialism, this thesis will argue six key points. It argues (1) that tool and source criticism are a requirement for the responsible use of digital tools in historical research, and it believes (2) that one of the best ways to perform this criticism is to build tools oneself. It will argue that (3) these tools need to be used as extensions of historical methods, not as replacements for them. It makes recommendations on the limits and best use cases for the two tools it develops, (4) topic modelling and (5) visualisation. Finally, it will make the point that (6) all tools are inherently limited, and that the best results can be gained from using multiple tools on multiple sources. These points will be argued throughout the following five chapters.

The first chapter of this thesis introduces the scholarly background to the project. It discusses the state of the Digital Humanities as a field, and how it has been shaping – and has been shaped by – the historical discipline. It argues that there exists a divide between the digital humanities and the wider historical community, in which both sides are wary of the other, which is in large part the result of many prior attempts to 'modernise' or 'rationalise' historiography by making it more scientific. It focusses in particular on the state of the art for the use of topic models to explore historical questions. It critically discusses the strengths

and weaknesses of these projects, in order to provide an introduction to the problems that are addressed by subsequent chapters of the thesis. Additionally, it will discuss the state of visualisation research, which is relevant for the development of the spatial visualisation tool. This chapter serves to bridge the divides between the three disciplines of history, computer science, and DH this project intersects with.

The second chapter discusses the material background of the thesis, providing an in-depth exploration of the newspaper archive, from its almost accidental conception to the deliberate choices made in the digitisation process. It follows the history of the articles this thesis draws on from their conception in the mind of a Victorian journalist to the moment they were stored on the project's server. For each of these stages, it discusses the implications for this project, and any others that draw from the same archive. It argues that at the point at which this project received the newspaper data it already carried the seeds within it that shape the research; by its very nature as a historical artefact it imposes constraints.

The third chapter discusses the topic modelling tool built by this project. It makes the case for Latent Dirichlet Allocation as an appropriate form of topic modelling to use on this archive.[27] It will provide a critical introduction to the technique, including its underlying statistical assumptions. It will then discuss the way the tool has implemented the LDA algorithm to work for its research questions and on its archival data. A large part of the work in this chapter is methodological. Throughout the discussion of the capabilities and operation of the

[27] Blei, Ng, and Jordan, 'Latent Dirichlet Allocation'.

tool it will continuously argue that, if the object of developing this tool is historical study, it has to shadow historical methodology, including being able to deal with uncertainty and levels of truth.[28] This also requires it to find a way to stay as close to the source material as possible. In presenting the way in which its topic modelling application works, it generates new insights into the methodological considerations that are involved in using these tools as parts of historical research. Additionally, it will investigate if the design and production of a tool in this way is feasible for an individual researcher.

Chapter four will present the methodological and theoretical basis for the visualisation of spatial patterns within newspapers, the methods this thesis has developed for visualising the location in which a subset of articles feature in a newspaper, and the process for interpreting the resulting heatmaps. It will show that there exists theoretical grounding for the visualisations it developed, in the form of earlier space-measurement analysis methods and the existence of limited case-studies by historians on individual or smaller samples of newspapers. It will provide an exploration of the methodology behind its spatial visualisations and the way it relies on the re-use of data intended for word- and article highlighting in the archive search interface. The fact that it is forced to use data in a different way than was intended means it has to make several assumptions about the relationship between the recorded page- and coordinate values. Next, in its discussion of the interpretation of the visualisations, it will explore the ancillary tools that are needed

[28] Nick Tosh, 'Science, Truth and History, Part I. Historiography, Relativism and the Sociology of Scientific Knowledge', *Studies in History and Philosophy of Science Part A*, 37.4 (2006), 675–701 <https://doi.org/10.1016/j.shpsa.2006.09.004>; John Tosh, *The Pursuit of History: Aims, Methods and New Directions in the Study of History*, 6th edn (London and New York: Routledge, 2015).

to understand the resulting visualisations, such as a general title-specific map of content. The chapter will demonstrate the validity of researching the non-textual aspects of a newspaper, and will present the implementation of article placement mapping developed by this project. It will argue that the visualisation of article placement allows for the exploration of historic trends in article placement design and the study of article layout, as well as allowing the spatial context of texts to re-emerge.

The final chapter has two main goals. The first is to show how the topic modelling and visualisation tools developed by this thesis can be integrated into the praxis of history. It argues this through example: by applying the tools to various case studies, it demonstrates the use cases for these tools, as well as their limits. The second goal of the chapter is to make a contribution to the discussion of banal imperialism, by testing the validity of MacKenzie's theories that imperialism was widespread in Victorian popular culture.[29] It will explore this theory through some of the facets Paul Ward has identified as contributing to a British national identity: Gender, Politics, and Leisure and Entertainment. It will supplement this with a study of the military, which is theorised by Colley to make a significant contribution.[30] It will also consider the economic relationship between Britain and its empire, which has been widely theorised.[31]

[29] *Imperialism and Popular Culture*, ed. by John M. MacKenzie (Manchester University Press, 1986).
[30] Linda Colley, *Britons: Forging the Nation 1707-1837*, 2nd edn (New Haven and London: Yale University Press, 2014).
[31] See from: *The Cambridge Economic History of Britain*, ed. by Roderick Floud and Paul Johnson, 2 vols (Cambridge: Cambridge University Press, 2014); Nuala Zahedieh, 'Overseas Trade and Empire', I, 392–420; Kevin Hjortshøf O'Rourke, 'From Empire to Europe: Britain and the World Economy', II, 60–94; Timothy J. Hatton, 'Population, Migration, and Labour Supply: Great Britain 1871-2011', II, 95–121.

Earlier, this thesis mentioned the metaphor of the sculptor. These chapters all fit within that schema. We study the material carefully (chapter 2), as it constrains the way it can be shaped: just as rough granite forms different shapes than smooth marble, so to an archive of hand-transcribed letters offers different opportunities than a machine-scanned collection of books. We construct and study our tools (chapters 3 and 4), as these influence the work we can do, and how we do it: while a fine chisel works well for details, it is inefficient for cutting rough shapes; topic models may work well on large datasets, but be of limited use on small ones. Based on these factors, material and tools, we craft our sculpture (chapter 5), which needs to fulfil all sorts of aesthetic and methodologic demands and will be judged on its own merit. But before all that, we have to study the context in which we work. All sculptors are shaped by the time during which they work, looking over the hedge to what other sculptors, painters and society as a whole are doing. Thus, the first chapter will look at the scholarly context of this thesis.

Chapter 1: Scholarly Background

Like so many witty sayings in the English language, the observation that 'there are no such things as new ideas' is often attributed to Mark Twain, although it is doubtful he ever wrote those words. Yet it has endured because the spirit of the saying holds true: every step in human endeavour involves building on the work of those that came before. Every sculptor is inevitably inspired to some extent by the artists that went before them, so it is appropriate and self-evident to discuss the scholarly background of this thesis. This takes the form of firmly rooting this thesis in the historiography and theory surrounding the interplay between computerised methods and writing history, which means any discussion on historiography will also have to take into account a significant portion of Digital Humanities theory. This, in turn, engages with several deeper debates about the nature of knowledge and the epistemological requirements of truth. These are the questions about the way tools have been used in the past, on different blocks of marble, to craft sculptures.

As this thesis sits at the intersection between several fields of study, it needs to clearly position itself in relation to the scholarly debates with more care than a non-interdisciplinary history PhD thesis. These will be the three fields of History, Computer Science, and the Digital Humanities. This chapter will first argue its historiographical validity, and its position in the two relevant debates. First, it will discuss the requirement that historical research needs transparency in its sources and methods, and that this goes double for projects using digital (or digitised)

sources and digital methods. After debating the nature of that transparency, it will continue discussing the way digital historical research projects have to navigate the divergent epistemological traditions of their method and subject matter. If we return to the metaphor of the sculptor, the historiographic debate corresponds to the collective practices of sculptors, regardless of their choice of medium. Next, this chapter will discuss this project's use of particular tools and techniques from computer science, and will argue for its choice of LDA as the analytical framework. It will also discuss some theory surrounding the use of visualisations in humanities research, and explore examples of the type of visualisation it intends to use. Metaphorically, these are the different chisels, which need to be discussed because the kind of chisel used affects the kind of sculpture one can make. Lastly, this chapter will argue the DH credentials of this project, and frame it in the debates surrounding the field's boundaries. While the metaphor begins to crack somewhat at this point, the digital humanities community can be seen as the stonemasons, who use the same kind of materials and techniques, but for a different purpose to a sculptor.

Because this thesis seeks to speak to readers from this wide range of disciplines, it can only make minimal assumptions to prior knowledge. Thus, it will occasionally cover ground that may be self-evident or basic knowledge for one of these fields, for the benefit of readers based in the others. For example, it offers a basic introduction to historical theory for the benefit of those from a computing background. Additionally, sometimes these fields intersect, and where appropriate this chapter will refer ahead to discussions later in the chapter.

The Sculptors Guild: Historiographical context

As this thesis has decided to investigate a historical phenomenon, namely, British imperial identity in the nineteenth century, and because this has been done using new and digital methods, it first needs to examine the historiographical validity of these techniques. In essence, as this thesis seeks to make claims on methodology and process, it needs to position itself with regards to historiography. It needs to address two main points. First, it needs to discuss the problem of transparency as the cornerstone of historical research in the context of 'black box' computational methods; next, the underlying tension that inevitably exits between the innovators and traditionalists, and the different epistemological traditions that both groups stand in. This discussion argues that there have been various attempts at incorporating the digital into historiography, but that these have failed to be adopted in ways that prompt a substantial re-thinking of methodological praxis beyond the world of quantitative historians.[1] That is not to say that digital history does not exist or is unimportant, but that it needs to make large strides to go from a new and experimental aspect of the discipline to a mainstream method of inquiry.[2] It argues that history, as a discipline, is more cautious about adopting methodological innovations than others. This is largely because academic history's claim to being a science is supported entirely by the merits and robustness of its

[1] Ernst Breisach, *Historiography: Ancient, Medieval, and Modern*, 3rd edn (Chicago & London: University of Chicago Press, 2007), pp. 370–72.

[2] Jane Winters and Steve F. Anderson, 'Digital History', in *Debating New Approaches to History*, ed. by Marek Tamm and Peter Burke (London: Bloomsbury, 2018), pp. 277–300 (pp. 288–90); Arguing with Digital History Working Group, *Digital History and Argument* (Roy Rosenzweig Center for History and New Media, 13 November 2017), p. 2 <https://rrchnm.org/wordpress/wp-content/uploads/2017/11/digital-history-and-argument.RRCHNM.pdf>.

method.[3] Nonetheless, this hesitance is in some cases being offset by increasing pressures on current generations of historians to produce more research in a shorter amount of time – and thus adopting digital methods and archives as time-savers.[4]

However, the methods that these historians adopted came with their own issues, as they were often adaptations or implementations of methods developed outside of historiography for linguists or for the purposes of information retrieval. This means the tools and methods need to be used while providing a firm explanation as to how they work and what they do for historical sources. Without showing the ins and outs of the methods, we create a what is termed a 'black box'; a mysterious device into which we input data and extract results without truly understanding the processes involved. In such a case, readers that are not intimately familiar with the way the method or tool works will be unable to criticise its findings and judge its value; after all, they do not fully understand how they were arrived at. Here lies the seed of the major issue this thesis has found in the application of digital tools to historical sources and the way in which the results from these investigations have been communicated to the broader community of historians. For those potential users that were not experts in the fields these tools

[3] This statement has some linguistic sensitivities to it. As a Historian trained in the Dutch and German tradition, in Belgium, I have always considered a 'science' to mean a field that is studied in academic context and with an academic rigour; in Dutch and German, there are no objections to speaking of *Historische Wetenschap* or *Geschichtswissenschaft* (litt. 'Historical Science') when discussing the academic practice of History. By contrast, in English the Natural Sciences are the only sciences. When I talk about History as a science, I mean it as a statement on its claim to epistemological soundness and discovery of truth. For more on these discussions see: Violet Soen, *Inleiding in de Historische Wetenschap*, (Leuven: ACCO, 2011), pp. 9-11; and Breisach, pp.272-290.

[4] Bingham, pp. 227–29; Lena Roland and David Bawden, 'The Future of History: Investigating the Preservation of Information in the Digital Age', *Library & Information History*, 28.3 (2012), 220–36 (pp. 228–31) <https://doi.org/10.1179/1758348912Z.00000000017>.

came from, the results would be distrusted, as they cannot judge the worth of the outcome or critique the underlying process.[5] For those who do have this knowledge, this sense of unease doesn't disappear, as they are well aware of the various ways in which the results may be skewed by all the (often undisclosed) factors involved in the analysis.[6] Thus, in order to explore the ways in which other researchers have dealt with, or in some cases have failed to address, these questions, it is useful to explore other sculptors' work.

Before this can take place, however, there is the need to determine what this thesis means by 'historiography', as there exists a subtle difference of meaning between the Dutch and German understandings of the term compared to the English. This needs to be discussed before all else, to ensure that any readers, whichever tradition they may come from, are on the same page with regards to terminology. Both are concerned with the way history has been written by previous practitioners. Yet the English interpretation of the term is slightly more concerned with the praxis of the historians of the past, seeking to describe how the historians of yesteryear wrote their histories, compared to the continental interpretation.[7] Amongst German, French and Dutch historians, there is more concern with the

[5] Fred Gibbs and Trevor Owens, 'Building Better Digital Humanities Tools', *DH Quarterly*, 6.2 (2012); Underwood; Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (World Scientific Publishing Company, 2015); Jennifer Guiliano, 'Toward a Praxis of Critical Digital Sport History', *Journal of Sport History*, 44.2 (2017), 146–59 <https://doi.org/10.5406/jsporthistory.44.2.0146>; Arguing with Digital History Working Group; Rik Hoekstra, 'Data Scopes for Digital History Research', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2, 2018, 1–16 <https://doi.org/10.1080/01615440.2018.1484676>.

[6] Mark Hall, 'Opportunities and Risks in Digital Humanities Research'. Pre-print 2020.

[7] Stefan Berger, Kevin Passmore, and Heiko Feldner, *Writing History: Theory & Practice*, Writing History, 2nd ed. (London: Bloomsbury Academic, 2010), pp. xi–xii <http://capitadiscovery.co.uk/edgehill/items/71829> [accessed 5 March 2018].

theoretical family a historian belonged to, and how their theoretical foundations underpin their methods.[8]

Compare, for example, the coverage of Maurist writing in two comparable historiographic textbooks: Ernst Breisach's *Historiography: Ancient, Medieval, and Modern* and Reginald de Schryver's *Historiografie: Vijfentwintig Eeuwen Geschiedschrijving van West-Europa*. Breisach, writing in the Anglo-Saxon tradition, introduces the group by stating "Unlike the Bollandists, the Maurists had only a loose work-program centering (sic) on a new history of the Benedictine order, … , and critical editions of some medieval works." He then discusses individual authors of the school in terms of their main works, while leaving their end open.[9] De Schryver writes from the Continental tradition, and by contrast, produces an extensive positioning of the social and academic context in which the Maurists operated, and the form this group collected under one moniker took: "The Maurists were not an institution; they were a collection and collaboration of learned individuals who shared initially a strong religious thesis, later a passion for erudition in its own sake. Their academic insights are of such quality that we may consider them as 'modern' historians." He subsequently discusses the two most notable authors of the school in detail, discussing what attributes they shared and how they were prototypical for the Maurists, and ends his discussion by covering their fall and successor organisations.[10]

[8] Reginald De Schryver, *Historiografie: Vijfentwintig Eeuwen Geschiedschrijving van West-Europa*, Ancorae, 8, 3rd edn (Leuven: Universitaire Pers Leuven, 1997), pp. 11–13.
[9] Breisach, pp. 194–95.
[10] De Schryver, pp. 240–43.

In this thesis, as a historian shaped by the continental tradition, I will be maintaining the continental definition. This means when discussing the historiographical context of the thesis I will focus more heavily on prototypical examples of schools of thought, rather than discussing any scholars' individual contributions. This leads to a discussion of the very foundation of historical scholarship as an academic field.

History, at least academic history, is practiced as an empirical study of the human experience of the past. This is far from a self-evident position, as by definition, the past is unavailable to be directly observed.[11] Why did the historical field take this position? To answer this, we need to look at the context in which history sought to be accepted as a field of study. When it first arrived on the scene in earnest during the nineteenth century, History as an academic discipline had to pay a fee to be taken seriously as an academic field: to adopt the tenets of empiricist science as was the fashion of the time. However, as the past is by definition unavailable for impartial observation, this leads to the question of how an academic historian should interpret the observations that they can access. One of the leading figures in finding an answer to this problem was the German scholar Leopold von Ranke.

While von Ranke is often presented as a solitary figure in laying the groundwork for academic history, in reality he was building upon the theories and methods developed by others, mainly the Göttingen school that had formed around textual scholars describing the wider context of the material they studied,

[11] Berger, Passmore, and Feldner, pp. 13–15.

and which was strengthened by Georg Hegel's historical philosophy in which each culture contributed its own unique advancements to mankind.[12] Von Ranke was inspired by these two insights when he wrote possibly the most-quoted words in modern historiography: "Man hat der Historie das Amt, die Vergangenheit zu richten, die Mitwelt zum Nutzen zukünftiger Jahre zu belehren, beigemessen: so hoher Ämter unterwindet sich gegenwärticher Versuch nicht: er will bloß zeigen, wie es eigentlich gewesen".[13] Von Ranke distanced himself from previous generations of writing about the past by developing an elaborate methodology to establish the historian as a neutral observer, based on classical literary studies, applying its maxim of strict criticism of a source's trustworthiness and a strong understanding of its context of creation and survival. The historian should aim to be as invisible as possible in this process, and not let their own views and biases reflect upon the past. Their aim was to produce 'pure' knowledge.

However, von Ranke was not a proto-positivist, seeking to use the objective building-blocks of knowledge he distilled from his sources to build a grand theoretical framework of human progress like Hegel had envisioned. Instead von Ranke was strongly influenced by the metaphysical philosophy of Kant: historical fact could not be used to induce increasingly general and abstract concepts, but it operated akin to Kantian 'ideas' – eternal and metaphysical forces which manifested themselves partially and temporarily in worldly phenomena. He disagreed with Hegel's belief in constant and inevitable progress and advancement;

[12] Breisach, pp. 217-219;231-232.

[13] History has been given the duty of making sense of the past, studying the world around it for the benefit of future years: such a high duty may not be interfered with by current ideas: it just needs to show what it actually was like. Leopold von Ranke, *Geschichte Der Romanische Und Germanische Völker von 1494 Bis 1514*, (Leipzig, 1885), Pp. VII., in *Inleiding Tot de Historische Wetenschap*, by Violet Soen (Leuven: Acco, 2011), p. 10.

"jede Epoche ist unmittelbar zu Gott", von Ranke wrote: each period is equal unto God, and therefore should be judged only on its own merits and against its own norms.[14] These ideas coalesced into a particular empirical and methodological form as historicism.

Von Ranke was hugely influential in the professionalisation of History as a discipline, particularly on the European mainland. However, from English and to an extent French historiography, the desire to create larger, overarching and structured narratives continued. As the natural sciences moved from understanding phenomena via observation to integrating these observations into ever more complex theories and models that could be used to predict as well as understand, some historians wished to step away from the bounds of periodisation in order to keep pace with the sciences. This desire is what led to more positivist system-builders of the late nineteenth and early twentieth century searching for systems and laws governing the past, such as H. T. Buckle and the second-generation Annales historians like Fernand Braudel. Their work eventually developed into its own particular brand of historiography, founded more along the quest for patterns and systems in human history based on a quantitative analysis.[15]

This division between historicist and positivist scholars caused a tension within academic history that shapes not only its own methodological processes, but also the way it collaborates with other disciplines. In this way, it serves as an example of C.P. Snow's 'Two Worlds' thesis. This divide between the two strands of epistemology in academia was addressed during the 1959 annual Rede lecture in

[14] De Schryver, p. 293; Breisach, pp. 233–34.
[15] De Schryver, pp. 337–42; 370–72.

Cambridge. C.P. Snow's lecture on *The Two Cultures and the Scientific Revolution* is relevant today as it was fifty years ago.[16] It discussed the difference in culture between the 'literary intellectuals' of the humanities and arts, and the natural scientist, between whom Snow saw a profound mutual suspicion and incomprehension. In turn, Snow saw in this divide a fundamental obstacle to the successful use of technology to address the ills of society, with the key difference being the way the two fields dealt with uncertainty. In the sciences, uncertainty must be minimised; in the humanities, it is celebrated. Even to this day, these questions loom over fields that intersect with history, such as Statistics or the Digital Humanities. One notable case in which this division expressed itself was the controversy on cliometrics, when computerised methods were considered for historical research without sufficient embedding in historical epistemology.

During the 1960s and 70s, the first attempts to introduce computers in mainstream historiography, championed by a school of economic historians who practiced cliometry, ended in a failure. Led by Robert Fogel and Stanley Engerman, cliometry attempted to introduce a quantitative turn to the writing of history, most notoriously in their 1974 work *Time on the Cross: The Economics of American Negro Slavery*.[17] Relying on thousands of data points relating to slavery, including census data and tax records, Fogel and Engerman set out to prove that slavery was an economically sensible and profitable system, and that the standard of living for black slaves was comparable to or better than that of a northern labourer. That

[16] C. P. Snow, *The Two Cultures*, 4th edn (Cambridge: Cambridge University Press, 2012).
[17] Robert W. Fogel and Stanley L. Engerman, *Time on the Cross: The Economics of American Negro Slavery* (London and New York: W. W. Norton, 1995).

research question, in and of itself, would have raised eyebrows amongst any American: Martin Luther King has been assassinated only five years prior, and the riotous summer of '67 was still fresh in people's memory. To the historian of the day, sensitive to politics of power in the aftermath of decolonisation, their thesis, that slavery was more benign and profitable than had been assumed, was an anathema. By discarding that which could not be quantified, the cliometricians could not convince historians that their methods were of any value to historiography, despite their vehement insistence that computerised methods were superior.[18]

In fact, part of the opposition to their work may have gained the emotional strength that it did because of their claims of educational and methodological superiority.[19] There was also a political aspect to this opposition, as many Southern commentators took *Time on the Cross* as a justification for their views on the institution of slavery and belief in 'liberal bias'.[20] Simultaneously, even amongst economic historians familiar with their statistical methods, their conclusions were drawn into question due to the implicit (erroneous) assumptions that had underlain their work. In R. Sutch's scathing rebuttal to the cliometric thesis, he writes:

> [I do not] object to the use of cliometric, economic or statistical procedures in the writing of history, but rather to the misuse of these techniques by Fogel and Engerman. … [I have conducted] a page-by-page and, in some cases, a line-by-line dissection of [some] major issues raised [in their book]. In each case I have found so many errors of computation or citation, data so selective or weak, and the presentation of results so distorted that I have

[18] F. Moura and P. Heitor, 'An Academic Parable: Robert W. Fogel's Raft', 2014; Herbert George Gutman, *Slavery and the Numbers Game: A Critique of Time on the Cross* (University of Illinois Press, 1975), p. 3.
[19] Richard Sutch, 'The Treatment Received by American Slaves: A Critical Review of the Evidence Presented in Time on the Cross', *Exploraions in Economic History*, 12.4 (1975), 335–438 (p. 337).
[20] Charles Crowe, 'Time on the Cross: The Historical Monograph as a Pop Event', *The History Teacher*, 9.4 (1976), 588–630 (pp. 599–600) <https://doi.org/10.2307/492099>.

> been forced to conclude that *Time on the Cross* is a failure. … It fails to establish the power or usefulness of cliometrics. [21]

In later editions of their work, Engerman and Fogel tried to pass off *Time on the Cross* as an "early report on preliminary findings", though this defence found little traction.[22] Since then, there have been many other attempts at using statistical methods to come to an objective truth about the past. Sometimes, these have generated valuable new insights, such as charting the rise of different modes of transport; but more often they have been strongly criticised by historians, such as a recent paper that received widespread public criticism after it statistically determined that people were happiest in the 1880s.[23]

The way the two sides of historiography interplayed during this 'cliometric crisis' can guide this thesis in its own use of digital techniques, as it highlights some red lines that the historical community considers impassable. First is the requirement of transparency: any newly proposed method must be understandable on at least a critical level by everyone, even those who have not studied it in depth. History as a discipline welcomes discussion and ambiguity, and thus the ability to discuss findings and the methods by which they have been observed from the sources needs to be transparent and open to debate. Second, this appreciation of ambiguity needs to be extended to the conclusions that these computer-centric methods draw about sources and about the past. Because the epistemological

---

[21] Sutch, pp. 338–39.

[22] Fogel and Engerman, p. 265; Paul A. David and others, *Reconing with Slavery* (Oxford University Press, 1985), pp. viii–ix; Moura and Heitor.

[23] Thomas Lansdall-Welfare and others, 'Content Analysis of 150 Years of British Periodicals', *Proceedings of the National Academy of Sciences*, 114.4 (2017), E457–65 <https://doi.org/10.1073/pnas.1606380114>; Thomas T. Hills and others, 'Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books', *Nature Human Behaviour*, 3.12 (2019), 1271–75 <https://doi.org/10.1038/s41562-019-0750-z>.

foundation of history as a discipline and the epistemological foundations of statistical or algorithmic methods are very different, this is vitally important, as a purely mathematical proof cannot stand on its own in a human context. As computers are not well placed to deal with ambiguity, not only is human mediation of the results necessary, but it also needs to carry with it a humility that the results it produced are not final. Essentially, it needs to be computer-assisted history instead of computerised history.

Allowing for ambiguity and contention is particularly important for this research project, as the area this project will apply its tools to, imperialism in Britain, is a contentious subject. There are multiple diverging readings on the spread and depth of imperial sentiment in Britain during the second half of the nineteenth century.[24] Some authors, such as John MacKenzie and the Manchester School see the prevalence of imperial symbols as a sign of broad public support for the empire; others, such as Bernard Porter look at this as an evidence of imperial propaganda trying and failing to move the masses. These diametrically opposed readings have emerged from the same historical sources, and no tool will be able to produce an objective, non-ambiguous answer when asked to analyse these documents. To add another layer of ambiguity, many of the frameworks these readings offer to understand the results are themselves subjective. They have been fostered by debate and discussion of ambiguous texts, such as Edward Saïd's readings of Orientalist literature. He relied on the same textual material that other

[24] See for example: MacKenzie, *Propaganda and Empire*; Bernard Porter, *The Absentminded Imperialists: Empire, Society and Culture* (Oxford: Oxford University Press, 2006); Jan Morris, *Pax Britannica: The Climax of an Empire*, Pax Britannica, 2nd edn, 3 vols (London: Faber & Faber, 2012), II; John Darwin, *Unfinished Empire: The Global Expansion of Britain* (London: Penguin Books, 2013).

scholars had used, but by reading it from a different perspective was able to develop wholly new insights; an algorithmic process would find it difficult to do this.[25]

This is also the case for the theoretical framework this thesis intends to use, the theory of Banal Nationalism.[26] This adds a level of semantic ambiguity: after all how can banality be measured in an objective fashion? A mathematical answer might focus on the pervasiveness of a phenomenon and consider anything that appears in more than a given percentage of the archive to be banal. While technically true, such an approach on an archive of newspaper articles between 2015 and 2020 would likely find 'Brexit' to be a banal concept, which is unlikely to reflect how readers experienced it. After all, the defining features of a banal phenomenon lie not just in its (potentially quantifiable) pervasiveness, but in the murkier and more qualitative aspects of its meaning and (in)significance. This realisation informs the use of the developed tools (topic modelling and spatial visualisation) in the case studies, and serves as a warning of overreliance on computational evidence.

Choosing the Chisels: LDA and Visualisation

Having discussed the historiographical context in which this thesis sits, and having covered the general epistemological and methodological scholarship, we now have to discuss the specific digital methods it intends to use. In the analogy of the sculptor, the choice of chisel is crucial for the final result. For the purposes of this

[25] Edward W. Said, *Orientalism* (London: Penguin, 1991).
[26] Billig, *Banal Nationalism*.

thesis, one of the two key 'chisels' is topic modelling; a technique which clusters similar texts together. For the topic modelling this thesis chooses to use the LDA algorithm developed by David Blei. This section argues that this was an appropriate method to use compared to various other techniques for generating a similar classification. In order to do this, it will discuss scholarship on these kinds of tools, beginning with the 1987 LSI algorithm. It will also discuss how other scholars have used similar tools. The goal of this section is not to provide an in-depth discussion of the operation of LDA, but to give the context in which it originated, and to provide a basic introduction to the technique for readers who are unfamiliar with topic modelling. A detailed discussion of the processes that underpin LDA will take place in chapter 3, during the implementation of these processes in the actual tool. Finally, it concludes by examining the second digital 'chisel' that this thesis seeks to use: visualisation.

Before this thesis can discuss the specifics of the LDA algorithm and its pedigree, it pays to expand somewhat on topic modelling on a conceptual level. As a tool, topic models are designed to help us identify patterns in large bodies of text that would be too large to read manually. At the most basic level, a topic model considers that there exist topics made up of words, and each of these words has a probability attached reflecting its importance or relevance to the topic. Each document has a number of topics attached, which each have a probability related to how much they contribute to the document. Each word in the document, then, is selected based on its weight within each topic and the weight of each topic for the overall text. All these factors are weighed against each other to come to a topic

model that captures the topic-word and document-topic relationships.[27] In concrete terms, words that frequently appear together in a text will end up in the same topic. For example, if 'railway' and 'timetable' frequently appear together, these two will be placed in the same topic. They may be joined by 'investors', if enough other uses of 'railway' are in conjunction with 'investors' – even if 'investors' and 'timetable' never appear together.

This epistemological foundation of topic models — the idea that the context of a word's appearance determines its meaning — is much older, and has roots in the work of the German philosopher Ludwig Wittgenstein, though only half a century after his death do we actually possess computers powerful enough to use his linguistic insights as the basis for textual analysis. In his 1953 work *Philosophical Investigations*, he discussed the philosophy of language. He argued that the meaning of a word is not immutably fixed, and that one can only know a word's meaning by observing how they are used at any one time.[28] Wittgenstein's observations are considered to have laid the foundations of modern corpus linguistics.[29] This focus on observing the word in its context fits with Wittgenstein's wider empiricist position, but topic modelling reframes it by using these observations to generate a predictive-descriptive model: it bases its

[27] Mark Steyvers and Tom Griffiths, 'Probabilistic Topic Models', *Handbook of Latent Semantic Analysis*, 427.7 (2007), 424–440; David Blei, *Prof. David Blei - Probabilistic Topic Models and User Behavior* (Edinburgh, 2017) <https://www.youtube.com/watch?v=FkckgwMHP2s&t=1086s> [accessed 24 February 2020]. Hall, 'Opportunities and Risks', pp. 7.

[28] Ludwig Wittgenstein, *Philosophical investigations*, 2nd ed. (Oxford: Blackwell, 1958) <http://capitadiscovery.co.uk/edgehill/items/14404> [accessed 23 January 2019].

[29] *Philosophy of Language and Linguistics: The Legacy of Frege, Russell, and Wittgenstein*, ed. by Piotr Stalmaszczyk, Philosophische Analyse (Boston, [Massachusetts]: De Gruyter, 2014), LIII <https://ebookcentral.proquest.com/lib/edgehill/detail.action?docID=1377128> [accessed 23 January 2019].

probabilities on the observed texts (descriptive), and can use these to predict which topic a new document belongs to.

One of the first of these applications was Latent Semantic Indexing (LSI). This algorithm came about in 1987 as the solution to a particular problem: how do we serve a user searching a collection of texts the most appropriate results for their query? In other words: how do we build the best search engine? Relying on a 1975 mathematical proof that showed all text can be represented as a collection of vectors, and that a comparison of two texts can thus be expressed as a comparison of vectors, LSI was intended to address this problem of building better search engines.[30] What set LSI apart from a simple text query was that it not only searched for a keyword, but automatically detected words that co-occurred with the keyword, as well as its contexts of use, to return a more focussed list of search results. [31] This was a significant advantage over older query-type interfaces, and LSI was quickly used in a multitude of search applications, ranging from online search engines to library catalogues.[32]

---

[30] G. Salton, A. Wong, and C. S. Yang, 'A Vector Space Model for Automatic Indexing', *Commun. ACM*, 18.11 (1975), 613–620 <https://doi.org/10.1145/361219.361220>.

[31] See for example patents on the technology: Scott C. Deerwester and others, 'Computer Information Retrieval Using Latent Semantic Structure', 1989; Thomas K. Landauer and Michael L. Littman, 'Computerized Cross-Language Document Retrieval Using Latent Semantic Indexing', 1994; Scott Deerwester and others, 'Indexing by Latent Semantic Analysis', *Journal of the American Society for Information Science*, 41.6 (1990), 391–407; On the underlying mathematical technique: G. H. Golub and C. Reinsch, 'Singular Value Decomposition and Least Squares Solutions', in *Linear Algebra*, ed. by J. H. Wilkinson, C. Reinsch, and F. L. Bauer, Handbook for Automatic Computation (Berlin, Heidelberg: Springer, 1971), pp. 134–51 <https://doi.org/10.1007/978-3-662-39778-7_10>; Charles F. Van Loan, 'Generalizing the Singular Value Decomposition', *SIAM Journal on Numerical Analysis*, 13.1 (1976), 76–83 <https://doi.org/10.1137/0713009>.

[32] Michael W. Berry, Susan T. Dumais, and Amy T. Shippy, *A Case Study of Latent Semantic Indexing* (Tennessee: University of Tennessee, 1995), p. 41 <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.1929&rep=rep1&type=pdf>; April Kontostathis and William M. Pottenger, 'A Framework for Understanding Latent Semantic Indexing (LSI) Performance', *Information Processing & Management*, Formal Methods for Information Retrieval, 42.1 (2006), 56–73 <https://doi.org/10.1016/j.ipm.2004.11.007>; Jing Kong, Alex Scott, and Georg M. Goerg,

However, LSI was found to be useful for more than just smarter searching. When not used as part of a search system, the tool for determining word use contexts and co-occurring terms could be used to classify which texts belonged together. When used in this context, it went by Latent Semantic Analysis, or LSA. Initially, LSA was the domain of a few researchers who had the technical skill and computational power to leverage, but as the 1990s progressed, the number of users increased.[33]

However, LSI/LSA had some flaws, which became more noticeable as more researchers sought to use it. The theoretical basis of the algorithm, mainly with regards to the choice of the number of topics, was found to be lacking – an issue which to this day has never been fully addressed. It is also difficult to provide reasoned mathematical interpretations of the topic-term and topic-document relationships that produces answers on a case-by-case basis. [34] To address these issues, a variant of LSA was created by Thomas Hofmann that had a stronger statistical foundation -Probabilistic Latent Semantic Analysis (pLSA).[35] The

'Improving Topic Clustering on Search Queries with Word Co-Occurrence and Bipartite Graph Co-Clustering', 2016.

[33] Landauer and Littman; Thomas K. Landauer, Peter W. Foltz, and Darrell Laham, 'An Introduction to Latent Semantic Analysis', *Discourse Processes*, 25.2–3 (1998), 259–84 <https://doi.org/10.1080/01638539809545028>; Charles R. Fletcher and Brian Linzie, 'Motive and Opportunity: Some Comments on LSA, HAL, KDC, and Principal Components', *Discourse Processes*, 25.2–3 (1998), 355–61 <https://doi.org/10.1080/01638539809545032>; Susan T. Dumais, 'Latent Semantic Analysis', *Annual Review of Information Science and Technology*, 38.1 (2004), 188–230 <https://doi.org/10.1002/aris.1440380105>; Jerome R Bellegarda, 'Statistical Language Model Adaptation: Review and Perspectives', *Speech Communication*, 42.1 (2004), 93–108 <https://doi.org/10.1016/j.specom.2003.08.002>; Steyvers and Griffiths.

[34] Chenggen Shi and Jie Lu, 'Choosing LSI Dimensions by Document Linear Association Analysis', in *Proceedings of the International Conference on Information and Knowledge Engineering*, 2003; Chenggen Shi and Jie Lu, 'A Text Mining Model by Using Weighting Technology', *AMCIS 2004 Proceedings*, 2004, 228; Dapeng Liu and Shaochun Xu, 'Challenges of Using LSI for Concept Location', in *Proceedings of the 45th Annual Southeast Regional Conference*, 2007, pp. 449–454.

[35] Thomas Hofmann, 'Probabilistic Latent Semantic Indexing', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99 (Berkeley, California, USA: Association for Computing Machinery, 1999), pp. 50–57 (p. 50) <https://doi.org/10.1145/312624.312649>.

'probabilistic' aspect of the method is that it assumes that each word contained in a document has an association with the topics that make up the corpus. This mixture of the words relative to the topic is described by a probabilistic distribution – certain topics have a higher chance of certain words appearing.[36] This addressed the key concern with LSA, and pLSA was swiftly adopted for topic modelling textual sources, including newspaper texts.[37] In some circumstances, pLSA remains competitive in accuracy with more modern topic modelling methods.[38]

However, pLSA had two main drawbacks. First, once the model was generated and the word-topic probabilities inferred, there was no way to add documents. As every corpus has its own word-topic probabilities, adding documents would mean calculating these probabilities anew for the collection of texts that are being added; however, pLSA does not account for a mechanism to determine which of these two word-topic probability distributions interact. Second, the number of parameters (topics) grows linearly with the amount of documents, which leads to overfitting – the state in which each document has its own topics which do not occur anywhere else in the corpus, making the modelling step lose its added value. There have been various adaptions of pLSA that sought to deal with these issues, including recursive (R)PLSA, incremental (I)PLSA,

---

[36] U. Naeem, A.-R. Tawil, and I.I. Kennedy, 'A Dynamic Segmentation Based Activity Discovery through Topic Modelling', in *IET International Conference on Technologies for Active and Assisted Living (TechAAL)* (presented at the IET International Conference on Technologies for Active and Assisted Living (TechAAL), London, UK: Institution of Engineering and Technology, 2015), p. 1 <https://doi.org/10.1049/ic.2015.0136>.

[37] Newman and Block; Berlin Chen, 'Latent Topic Modelling of Word Co-Occurence Information for Spoken Document Retrieval', in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3961–64 <https://doi.org/10.1109/ICASSP.2009.4960495>.

[38] Anna Potapenko and Konstantin Vorontsov, 'Robust PLSA Performs Better Than LDA', in *Advances in Information Retrieval*, ed. by Pavel Serdyukov and others, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer, 2013), pp. 784–87 <https://doi.org/10.1007/978-3-642-36973-5_84>.

folding-in (fold.) PLSA, quasi-Bayes PLSA, and Online (O)PLSA; none of which gained major traction.[39] In part, this was the result of most of these adaptations being closed-source code, which were not made available in easy-to-use packages and tools.

In response to these issues with pLSA and its various offshoots, David Blei, Andrew Ng, and Michael Jordan developed the Latent Dirichlet Allocation (LDA) algorithm.[40] The gist of LDA is that it randomly assigns words the probability of belonging to a topic, and then continuously revaluates these probabilities until they hold true based on the observed word co-occurrence statistics in an iterative process. Each iteration updates the word-topic assignments based on what it has 'learned' on its last pass through the documents in the collection, only stopping when either a given maximum number of iterations is reached, or when word-topic assignments remain unchanged. Increasing the number of modelling passes thus significantly influences the reliability of the final model.[41] LDA addresses the two key weaknesses of pLSA: it has a set and immutable number of topics, which it needs to assign words to at the beginning of the process; and it allows for the addition of documents to a model after generating, by adding them to the 'pile' of documents to draw frequency inference from. LDA became a standard of topic modelling swiftly, and has remained popular.[42] In part, this popularity may be

[39] Nikoletta K. Bassiou and Constantine L. Kotropoulos, 'Online PLSA: Batch Updating Techniques Including Out-of-Vocabulary Words', *IEEE Transactions on Neural Networks and Learning Systems*, 25.11 (2014), 1953–66 <https://doi.org/10.1109/TNNLS.2014.2299806>.

[40] Blei, Ng, and Jordan, 'Latent Dirichlet Allocation'.

[41] Blei.

[42] Erik Linstead and others, 'Mining Concepts from Code with Probabilistic Topic Models', in *Proceedings of the Twenty-Second IEEE/ACM International Conference on Automated Software Engineering - ASE '07* (presented at the the twenty-second IEEE/ACM international conference, Atlanta, Georgia, USA: ACM Press, 2007), p. 461 <https://doi.org/10.1145/1321631.1321709>; Jonathan Chang and others, 'Reading Tea Leaves: How Humans Interpret Topic Models', in *Advances in Neural Information Processing Systems*, 2009, pp. 288–

explained by the availability of LDA-tools such as MALLET and GenSim, which allow for simple use by researchers.[43] Topic models as a historical research tool remained fairly obscure until the mid- to late 2010s, when they underwent a major spike in popularity. Recently, a limited LDA topic modelling extension was even added to the *Gale Digital Labs* suite, integrating it into a commercial archive interface designed to be used by non-experts.[44]

Several other researchers have already used LDA extensively on historical sources. However, there are two distinct groups of users to be observed. First are the researchers that focus on the technical aspect of their work and use historical collections simply because they are an available body of material with a known context. One example of this kind of work is the paper by Yang, Torget and Mihalcea, using LDA on Texan newspapers from 1829 to 2008, which does not attempt to produce historically sound knowledge, but instead looks to experiment with modelling settings, parameters, and workflows. [45] This is reflected in its selection of historical periods to investigate; not only are these internally inconsistent (1865-1901; 1892; 1893; 1929-1930), they also offer little in the way of

296; Clare Llewellyn, Claire Grover, and Jon Oberlander, 'Summarizing Newspaper Comments', in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014 <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8098> [accessed 24 February 2020]; Clare Llewellyn, Claire Grover, and Jon Oberlander, 'Improving Topic Model Clustering of Newspaper Comments for Summarisation', in *Proceedings of the ACL 2016 Student Research Workshop* (presented at the Proceedings of the ACL 2016 Student Research Workshop, Berlin, Germany: Association for Computational Linguistics, 2016), pp. 43–50 <https://doi.org/10.18653/v1/P16-3007>; Quintus van Galen and Bob Nicholson, 'In Search of America: An Introduction to Topic Modelling Nineteenth-Century Newspaper Archives', *Digital Journalism*, 2018.

[43] Andrew K. McCallum, *MALLET: A Machine Learning for Language Toolkit* (University of Massachusetts Amherst, 2002) <http://mallet.cs.umass.edu>; Radim Rehurek and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, 2010, pp. 45–50.

[44] Gale-Cengage.

[45] Tze-I Yang, Andrew J. Torget, and Rada Mihalcea, 'Topic Modeling on Historical Newspapers', in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11 (Stroudsburg, PA, USA: Association for Computational Linguistics, 2011), pp. 96–104 <http://dl.acm.org/citation.cfm?id=2107636.2107649> [accessed 26 September 2017].

justification for their selection. The paper's focus on the workflow and the pre-processing of historic research is useful, but as it lacks grounding in historiography, the implications of its choices on the historical output are uncertain. It proposes the use of various pre-processing steps, such as a Dictionary-based OCR clean-up, dehyphenation, NER, and a stemmer. However, it fails to discuss the selection process of the dataset it runs on, in contrast with some of the other practical applications (notably Block). The actual conclusions it draws on history are almost non-existent, only bringing in their historian to determine whether their topics were meaningful. This same sense of technology taking centre stage permeates the project's follow-up article, 'Mapping Texts'.[46]

In the second category are those scholars that sought to do what this thesis does: find a way to employ topic models in a historically sound manner, often as an innovative solution to the overwhelming amount of material a historian has to contend with, and then use it as a way to answer historical research questions. Of this last group, this thesis discusses four examples: Newman and Block's ground-breaking project topic modelling the eighteenth-century *Pennsylvania Gazette*; Nelson's tools for analysing the *Richmond Dispatch*, which allow for a quick transit from distant to close reading; Miller's analysis of the meaning of 'bandit' in nineteenth-century Chinese sources by topic decomposition; and Giffard and van den Bos' paper on topic modelling the Dutch newspaper corpus for the development of national identity.

[46] Andrew J. Torget and others, 'Mapping Texts: Combining Text-Mining and Geo-Visualization to Unlock the Research Potential of Historical Newspapers', *University of North Texas Digital Library*, 2011 <https://pdfs.semanticscholar.org/4b40/d6b77b332214eefc7d1e79e15fbc2d86d86a.pdf> [accessed 26 September 2017].

Block's 'Doing more with Digitisation' blog post is a companion piece to Block & Newman's 'Probabilistic Decomposition of an Eighteenth-Century American Newspaper', published in 2006.[47] The latter is the first paper in which historical subject matter is investigated using topic modelling. The website takes a more historical angle than the technical journal article, and overall provides a good evaluation of the usefulness of the method: topic modelling is only a tool. It requires historians' knowledgeable input and analysis. Unfortunately, it is also a tool that requires not only access to the text (rather than page images) of documents but the cooperation of a computer scientist. It discusses the ways in which to analyse the historical rise and fall of topics within a corpus, although they seem to prefer using a probabilistic method over a topics histogram, which might have something to do with the computing times back when the article was written. The paper was the first to show that topic modelling is a potentially useful approach to take when questioning historical corpora, and that the topics can lead to unexpected insights, when properly supervised by a historian. However, their work was published in an information science journal, putting it in the periphery of historians' view, and their attempt at innovating historical methods passed by largely unnoticed in the historical community.[48]

The next attempt at incorporating topic models into the historical toolbox was the 'Mining the Dispatch' project; an online project from 2010 by R. Nelson to present his work on topic modelling the *Richmond Daily Dispatch*, including a clear

[47] Newman and Block.

[48] While the article has racked up an impressive 150 citations on Google Scholar by 1/2/2020, many of these are computer science papers showing the relevance of work to test or improve LDA. Citations from historians and humanities scholars only pick up from 2015 onwards.

exploration of the strengths and weaknesses of his approach. It shows a well-developed interface for interpreting the results using a simple case study, merging article frequency counts (topic histograms) and topic modelling.[49] The project provides the insight that the generated topics can be fed back into the corpus as queries, to discover which documents are most strongly correlated with a particular topic. It also illustrates the power of LDA on a clean and well-curated dataset, discovering highly specific topics. Like many other historians using topic models, Mining the Dispatch uses a single title as a source of its texts. This is the most methodologically sound way to use LDA, as it assumes the chance of a word being included in a document is equal for each – thus it would distinguish different writing styles as different topics. This is one place where this thesis seeks to make a methodological contribution to knowledge; it will assess the possibility of topic modelling multiple newspaper titles in a single model. While others have modelled multiple titles, the methodological implications have remained largely underexplored.[50]

Unfortunately, around the time Nelson produced his findings, the Digital Humanities, especially in the United States, were at the centre of a cultural and political row over ideology and funding. The Digital Humanities community were seen by some humanities researchers as part of a 'neoliberal push' into the left-leaning humanities faculties, in part because these programmes were the only ones left that still received funding in the post-crash austerity environment.[51] This

[49] Robert K. Nelson, 'Mining the Dispatch', 2010 <http://dsl.richmond.edu/dispatch/pages/intro> [accessed 26 September 2017].
[50] See, for example, H. Giffard and M. van den Bos' work below.
[51] Richard Grusin, 'The Dark Side of Digital Humanities: Dispatches from Two Recent Mla Conventions', *Differences*, 25.1 (2014), 79–92 <https://doi.org/10.1215/10407391-2420009>. Allegations later repeated by

reading of the Digital Humanities sees it as an enabler of casualisation of academic labour, and a means by which university management can impose 'measurable' targets on humanities departments. These debates masked the significance of the work, which was further hampered by its nature as an online resource rather than a journal article or monograph.[52]

Three years later, in 2013, I. M. Miller researched crime and unrest in eighteenth- and nineteenth-century China through topic modelling for the distant reading special issue of *Poetics*.[53] His paper provides an extremely valuable case study into the practical applicability of the method in historiography. Unlike most other users, he provides a good argument for the use of topic modelling to explore the Great Unread, while still acknowledging that it is not well-suited for exploring the meaning of single documents. It is, however, a great tool for exploring contentious words/definitions in large corpora, as it leaves the researcher free from a priori definitions of any topic – as is the risk with a keyword-based approach. His use of targeted close reading of the texts that are the best match for a topic in order to understand the nuances of that topic goes a long way towards bridging the methodological divide, as it provides a point of entrance for the average historian trying to value his findings. The article proves that topic models work as a tool of historical research, as it succeeds in identifying the topical differences between various meanings of crime, rebellion and banditry. Yet it was only one article,

Daniel Allington, Sarah Brouillette, and David Golumbia, 'Neoliberal Tools (and Archives): A Political History of Digital Humanities', *Los Angeles Review of Books* <https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/> [accessed 23 July 2019].

[52] The 43 citations of the work in Google Scholar as of 22/2/2020 are mainly from publications after 2016, when topic modelling began to rise in popularity.

[53] Ian Matthew Miller, 'Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach', *Poetics*, 41.6 (2013), 626–49 <https://doi.org/10.1016/j.poetic.2013.06.005>.

amongst a special issue aimed at those already using distant reading techniques, on a niche subject and using a specialised archive.[54]

The most recent attempt at wider adoption of topic models at the time of writing is H. Giffard and M. van den Bos' article 'Mining Public Discourse for Emerging Dutch Nationalism', which uses them to analyse the Dutch newspaper corpus *Delpher* for traditional markers of nationalism.[55] Their research uses a method for dealing with temporal and topical variation that this thesis will adopt: by making sub-corpora around a specific keyword and a specific date, and applying the topic modelling algorithm to those. This has resulted in the discovery that nineteenth-century Dutch newspapers lack the emotional nationalism that might be expected, but instead show an increase in the 'banal nationalism' theorised by Michael Billig.[56] However, their paper also illustrates the issue that many non-digital historians have with the Digital Humanities, when it admits that its core methodology is "poorly understood", and it makes no attempt to justify its research design.

As was already alluded to by Giffard and van den Bos, despite its popularity, LDA, and topic models more generally, are not uncontroversial amongst historians, and are not completely understood. Some of this criticism relates directly to the way LDA operates: as it discards many 'common words' that are shared between documents and makes word order irrelevant when reducing a text

[54] As of 1/2/2020, the paper was cited 16 times in a historical context on Google Scholar.
[55] Bos and Giffard.
[56] Billig, *Banal Nationalism.*

to a bag-of-words, which causes some concern for loss of meaning.[57] While topic models are very useful for classifying texts and providing a general overview of a corpus, the cohesion of the resulting topics is not always good.[58] But more fundamentally, algorithms, like LDA themselves are an issue in humanities scholarship, as they are unable to deal with the required level of ambiguity.[59] Algorithms are fundamentally biased towards providing an answer – and are generally incapable of modelling when they can't.[60] Added to this is the issue that it is often unclear what a model has 'learned': if we determine a topic to be letters from readers, do documents end up there because they are letters or because of some underlying factor in the data that we can't see? Several researchers have shown that minute changes to the data given to modelling and learning algorithms can considerably affect their outcomes.[61] Additionally, LDA is often not evaluated

[57] Alexandra Schofield and others, 'Understanding Text Pre-Processing for Latent Dirichlet Allocation', 2017 <https://www.cs.hmc.edu/~xanda/files/winlp2017.pdf> [accessed 10 February 2019].

[58] Ruidan He and others, 'An Unsupervised Neural Attention Model for Aspect Extraction', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (presented at the Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada: Association for Computational Linguistics, 2017), pp. 388–97 (p. 388) <https://doi.org/10.18653/v1/P17-1036>.

[59] D. Sculley and Bradley M. Pasanek, 'Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities', *Literary and Linguistic Computing*, 23.4 (2008), 409–24 <https://doi.org/10.1093/llc/fqn019>; Jennifer Edmond, 'Managing Uncertainty in the Humanities: Digital and Analogue Approaches', in *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18 (Salamanca, Spain: Association for Computing Machinery, 2018), pp. 840–844 <https://doi.org/10.1145/3284179.3284326>; Roberto Therón Sánchez, Antonio Losada Gómez, and others, 'Toward Supporting Decision-Making under Uncertainty in Digital Humanities with Progressive Visualization', in *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18 (Salamanca, Spain: Association for Computing Machinery, 2018), pp. 826–832 <https://doi.org/10.1145/3284179.3284323>; Roberto Therón Sánchez, Alejandro Benito Santos, and others, 'Towards an Uncertainty-Aware Visualization in the Digital Humanities', *Informatics*, 6.3 (2019), 31 <https://doi.org/10.3390/informatics6030031>.

[60] Hall, Mark, 'Opportunities and Risks in Digital Humanities Research', In Print (2020), pp. 13.

[61] Anish Athalye and others, 'Synthesizing Robust Adversarial Examples', *ArXiv:1707.07397 [Cs]*, 2018 <http://arxiv.org/abs/1707.07397> [accessed 25 February 2020]; Samuel G. Finlayson and others, 'Adversarial Attacks Against Medical Deep Learning Systems', *ArXiv:1804.05296 [Cs, Stat]*, 2019 <http://arxiv.org/abs/1804.05296> [accessed 25 February 2020]; Dimitris Tsipras and others, 'Robustness May Be at Odds with Accuracy', *ArXiv:1805.12152 [Cs, Stat]*, 2019 <http://arxiv.org/abs/1805.12152> [accessed 25 February 2020].

properly by humanities scholars, which leads to the possibility that random noise is presented as research results.[62]

Given these scholarly question marks next to LDA (and topic modelling as a whole), can this thesis even consider using it? I believe so, within certain constraints. Latent Dirichlet Allocation is a useful tool when used with the appropriate degree of epistemological scepticism; even its most vocal detractors accept that it does produce valid clustering of text, even if these may not be reproducible and inscrutable. We thus need to methodologically return LDA to where it came from: a search aid. The results of an LDA analysis cannot be taken as definitive answer, only as an indicator of what might be. Based on the criticism it has faced, LDA reminds me of one of the aspects of the Honda Point Disaster of 1923, in which seven U.S. Navy destroyers ran aground off the coast of California due to a failure of the dead reckoning method of navigation used.[63] The ships that foundered did not measure their speed directly, rather they derived it from the rotations of their screws; as the sea was rough, these broached the surface and fouled the reckoned speed.[64] LDA is similar, as it provides an estimate of the composition of a corpus, one that is good enough in most cases – but one should err on the side of caution when navigating by it to a research outcome, lest one runs aground.

---

[62] Nan Z. Da, 'The Computational Case against Computational Literary Studies', *Critical Inquiry*, 45.3 (2019), 601–39 <https://doi.org/10.1086/702594>.

[63] Dead reckoning is one of the simplest forms of navigation without using landmarks. The method relies on taking the course, time, and speed from the last known location and extrapolating the current position based on trigonometry. For example, a ship sailing due west from Liverpool at 12 miles an hour will reckon it can safely turn south past Holyhead (a distance of 66 miles) after 5.5 hours.

[64] Noah A. Trudeau, 'A Naval Tragedy's Chain of Errors', *Naval History Magazine*, 24.1 (2010) <https://www.usni.org/magazines/naval-history-magazine/2010/february/naval-tragedys-chain-errors>.

The second main contribution of this thesis centres on the use of visualisations to expose the spatial properties and patterns of the newspaper articles it investigates. Specifically, it visualises the density of article placement for specific subsets selected by keywords search in a heatmap. It is important, therefore, to discuss the epistemology of visualisation in general, and the prior uses of heatmaps in the humanities. A heatmap is a matrix of data, consisting of a rectangular tiling grid, with each tile shaded on a colour scale to represent the value of the corresponding element of the data matrix.[65] An example is included as Figure 1.1. Within the application of visualisation to humanities projects, Stefan Jänicke has identified the problem of the tightrope: while the visualisation scholar would like to make a theoretical contribution and develop a novel approach to visualising humanities data, the humanist wants to be able to understand the visualisation process to avoid a 'black box'.[66] This is especially important as visual information is often perceived to be more convincing than merely textual information, despite it representing a reduced and less nuanced truth.[67] As all visualisations are inherently ideological and subjective, we thus have a duty, as

[65] Leland Wilkinson and Michael Friendly, 'The History of the Cluster Heat Map', *The American Statistician*, 63.2 (2009), 179–84 (pp. 179–80) <https://doi.org/10.1198/tas.2009.0033>.
[66] Jänicke, p. 4; Bernhard Röhle and Theo Rieder, 'Digital Methods: Five Challenges', in *Understanding Digital Humanities*, ed. by David M. Berry (London: Palgrave Macmillan UK, 2012), pp. 67–84 (p. 76) <https://doi.org/10.1057/9780230371934_4>.
[67] Röhle and Rieder, p. 74; Hein van den Berg and others, 'A Philosophical Perspective on Visualization for Digital Humanities', in *3rd Workshop on Visualization for the Digital Humanities* (presented at the IEEE VIS 2018, Berlin, 2018), p. 4 <http://vis4dh.dbvis.de/papers/2018/A%20Philosophical%20Perspective%20on%20Visualization%20for%20Digital%20Humanities.pdf>.

humanities scholars, to be completely transparent in the description of how the visualisation was generated and what all aspects of it mean.[68]

	Left	Middle	Right
Top	22	45	84
Middle	23	61	37
Bottom	108	54	72

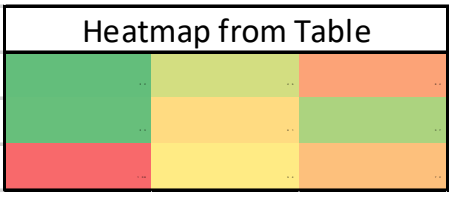


Figure 1.1 Example of a Heatmap. The values of the table left have been translated to colour shades in the image right. This example uses a green-yellow-red colour gradient.

We will now look at some other researchers that have relied on visualisations to present the stories of their data. Some, such as Natalie Houston, have visualised the metadata of texts themselves.[69] Thus far heatmaps have not often been used to visualise the physical location of a text on a page. There have, however, been projects that used them as a part of literary analysis. One key example of such use is Daniella Oelke et al (2012), who used heatmaps, amongst other visualisation techniques, to explore thirteen Swedish novels written between 1909 and 1930.[70] The visualisations produced by their project are effective at 'telling the story of the data', for example showing the patterns of appearance of certain (types of) characters, in a way that would not be possible without them. However, they also highlight the dangers of opaque visualisation practices. The graphs presented in their work lack features essential to understanding them, such as accurate (or even any) scales; One of the heatmaps has its colour bar marked

[68] Athir Mahmud and others, 'Teaching Students How (Not) to Lie, Manipulate, and Mislead with Information Visualization', in *Big Data Factories*, ed. by S. A. Matei, S. P. Goggins, and N. Jullien (Springer, 2017), pp. 101–114 (p. 103).
[69] Natalie M. Houston, 'Towards a Visual Analysis of Victorian Poetics', *Victorian Studies* 56.3 (2014), 498-510.
[70] Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm, 'Advanced Visual Analytics Methods for Literature Analysis', in *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '12 (Avignon, France: Association for Computational Linguistics, 2012), pp. 35–44.

simply with 'few' to 'many'.[71] These issues also plague other work by the same author.[72]

Others have used heatmaps as a way to chart the similarities between different texts. These visualisations place texts from different collections on the X- and Y-axis, while using a similarity metric to determine the shade of each square. Two examples of this can be found in the proceedings of the 2017 Japanese Association for Digital Humanities conference.[73] Both of these papers highlight two crucial issues with heatmap visualisation design: they rely strongly on colour, and the absence of colour makes it much more difficult to see the nuances in the graph; and the use of a single gradient, in the case of these images white-black, provides an additional interpretative difficulty.

Looking at the existing state of the art of heatmap visualisation, the key discovery is the relative lack of theoretical grounding of this type of graph. There exists little in the way of debate on when this is and isn't an appropriate type of visualisation to use. What is apparently present is a discussion more concerned with the praxis of the visualisation, such as the colours used. [74] The consensus that does exist however, is that visualisations need to 'tell the story of the data'. A

[71] Oelke, Kokkinakis, and Malm, fig. 2;3. (pp. 41-42)

[72] Daniel A. Keim and Daniela Oelke, 'Literature Fingerprinting: A New Method for Visual Literary Analysis', in *2007 IEEE Symposium on Visual Analytics Science and Technology* (presented at the 2007 IEEE Symposium on Visual Analytics Science and Technology, Sacramento, CA, USA: IEEE, 2007), pp. 115–22 <https://doi.org/10.1109/VAST.2007.4389004>.

[73] Peter Broadwell, Tomoko Bialock, and Hiroyuki Ikuurada, 'Macroscopic Exploration of Large Text and Image Collections via Similarity Heatmaps', in *Proceedings of the 7th Conference of Japanese Association for Digital Humanities* (presented at the JADH 2017, Doshisha, 2017), pp. 1–4; Tomoji Tabata, 'Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction', in *Proceedings of the 7th Conference of Japanese Association for Digital Humanities* (presented at the JADH 2017, Doshisha, 2017), pp. 73–78.

[74] David Borland and Russell M. Taylor, 'Rainbow Color Map (Still) Considered Harmful', *IEEE Computer Graphics and Applications*, 27.2 (2007), 14–17 <https://doi.org/10.1109/MCG.2007.323435>.

visualisation or graph needs to make or support an argument and show insight into the data that otherwise would not be gained. For this to be possible, visualisations need to be produced and presented with the greatest possible transparency on behalf of the creator.

Stoneworkers of the Wider World: The Digital Humanities

The previous sections have discussed the historic context in which this thesis operates, as well as the methods it intends to use and under what circumstances. Which begs the final question in this chapter: does the choice of LDA and visualisations as methods make this thesis a DH thesis? Is the use of a digitised archive and digital research infrastructure and methods enough to qualify as a digital humanities project? The answers to this question are tied up in one of the major debates within the Digital Humanities, surrounding the boundaries of the field and the importance of 'making'. It is true that DH is no longer the fledgling discipline that it was in the first decade of the 21st century, when it was still trying to define itself (and sailing under the flag of 'Humanities Computing').[75] Nor is the discipline trying to prove its validity and staying power, spreading exponentially across academic institutions, as it was in the early- and mid-2010's. Now, at the start of a new decade, DH has firmly established itself as a valid discipline, with its own apparatus of academia around it. However, the field still struggles with self-definition; ask a dozen attendees at a major DH conference what defines the field, and you will hear a dozen different answers.[76]

[75] Robert Scholes and Clifford Wulfman, 'Humanities Computing and Digital Humanities', *South Atlantic Review*, 73.4 (2008), 50–66.

[76] See for a discussion on this definition: Melissa Terras and others, *Defining Digital Humanities : A Reader* (Routledge, 2016) <https://doi.org/10.4324/9781315576251>; Sarah K. Sanders, 'Assessing the Missions

In short, the question is thus: where lies the boundary between those humanities researchers that use computers, and "the digital humanities researcher"?[77] As Stephen Ramsay put it: "Who's in and who's out."[78] The origins of this debate lie in the rapid expansion of the field in the 2000s. At this time, the field changed the name it had borne from the 1950s, Humanities Computing, to Digital Humanities.[79] Before then, the field aimed to be as inclusive as possible; the introduction to the first issue of *Computers and the Humanities* considers itself to be a full part of the humanities:

> We define humanities as broadly as possible. Our interests include literature of all times and countries, music, the visual arts, folklore, the non-mathematical aspects of linguistics, and all phases of the social sciences that stress the humane. When, for example, the archaeologist is concerned with fine arts of the past, when the sociologist studies the non-material facets of culture, when the linguist analyses poetry, we may define their intentions as humanistic; if they employ computers, we wish to encourage them and to learn from them.[80]

This inclusive position became increasingly controversial due to the growing accessibility and usability of computers and digital resources and the associated rapid expansion of the field of digital humanities in the 2010s. The fundamental distinction posed in the debate that followed was between the 'makers' and the 'thinkers'. The latter wished to see a continuation of the inclusive meaning of DH,

---

of Digital Humanities Centers', 2019; Dino Buzzetti, 'The Origins of Humanities Computing and the Digital Humanities Turn', *Humanist Studies & the Digital Age*, 6.1 (2019), 32–58.

[77] John Unsworth, 'What Is Humanities Computing and What Is Not?', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 35–48; Lincoln Mullen, 'Digital Humanities Is a Spectrum, or "We're All Digital Humanists Now"', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 237–38. Original Published in 2002.

[78] Ramsay, 'Who's In and Who's Out'.

[79] See for a brief history of Humanities Computing before 2010: Edward Vanhoutte, 'The Gates of Hell: History and the Definition of Digital | Humanities | Computing', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 119–56.

[80] 'Prospect', *Computers and the Humanities*, I.1 (1966), 1–2.

while the former proposed a much stricter definition of the field. A Digital Humanist had to be more than just a user of tools created by others, but be involved in the creation (or criticism) of their own tools, which required skills such as being able to code.

One of the influential voices on the side of the makers was Stephen Ramsay, who took an unequivocal stand for the process of creating within DH.[81] He posited that the inclusive nature of DH was diluting the field, and that it was endangering the field's future. Out of fear of seeming exclusionary, DH as a whole was refusing to engage with its disciplinary boundaries.

> [When talking about the definition of DH] we go further, and say that it doesn't really matter. Everyone is included. It's all about community and comity, collaboration and cooperation. But this, of course, is complete nonsense. Community and collaboration are undoubtedly signs of the spirit, but to say that disciplinary definition doesn't really matter is to eschew the hard reality of life in the modern academy. Digital humanities is not some airy Lyceum. It is a series of concrete instantiations: … . It might be more than these things, but it cannot not be these things.[82]

While Ramsay accepted the danger of an overly precise definition, and that it is also possible to have a field survive and even flourish through an internal schism, he warned against overgeneralising the Digital Humanities.

In Ramsay's view, the process of 'making' is integral to the process of the Digital Humanities generating knowledge. "Building is, for us, a new kind of hermeneutic – one that is quite a bit more radical than taking the traditional methods of humanistic inquiry and applying them to digital objects."[83] He, and

---

[81] Stephen Ramsay, 'On Building', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 243–45.
[82] Stephen Ramsay, 'Who's In and Who's Out', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 239–41 (p. 240).
[83] Ramsay, 'On Building', p. 244.

those of the 'maker'-school with him, held that the price of entry into the Digital Humanities as a separate field was to become a builder. Allowing for a wide range of activities to fall under that header, such as visualising, coding, marking up, hacking, map-making, the 'makers' conclude that the desire to have, at the end of the research process, a tangible product, is what sets DH apart.

In opposition to these 'makers' stood the 'thinkers', such as Geoffrey Rockwell and Mark Sample. Rockwell notes that in the period between 1950 and 2010, "Our biggest problem for years was getting anyone to come to meetings, especially graduate students. We traded stories about the lack of respect from the established disciplines and how we had sacrificed traditional careers to pursue computing", noting the irony in the desire to exclude those outside the field's boundaries.[84] As he and John Unsworth both note, the 'makers' seem to desire specific skills as markers for inclusion, but due to the fluidity of the technology in the field that those skills apply to, such a boundary line may be obsolete within a few years.[85] Rockwell observes a "residual fear … of theory", as in the hands of institutional mechanisms and funding organisations, it may lead to suppression.[86] Yet theory, and the inclusivity of non-makers that it requires, are what the field needs, as "the ideal would be to develop a space where the theoretical and the pragmatic can inform each other without participants needing to excel at both."[87] He proposes a Digital Humanities that is modelled on the role statistics plays for

[84] Geoffrey Rockwell, 'Inclusion in the Digital Humanities', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 247–53 (p. 248).
[85] Rockwell, p. 249; John Unsworth, 'The State of the Digital Humanities', 2010 <http://www.people.virginia.edu/~jmu2m/state.of.dh.DHSI.pdf>.
[86] Rockwell, pp. 249–50.
[87] Rockwell, p. 251.

the social sciences, where it is both an ancillary discipline as well as a field of study in its own right.[88] Sample, meanwhile approaches the debate from a standpoint of reproduction, rather than production (of either knowledge of tools). "the promise of the digital is not in the way it allows us to ask new questions because of digital tools or because of new methodologies made possible by those tools. The promise is in the way the digital reshapes the representation, sharing, and discussion of knowledge."[89] In other words, Sample argues that the division between builders and thinkers is a false one, as both are united in the desire to transform existing academic structures.[90]

This thesis finds itself more receptive to the 'makers' side of the argument. It agrees with Ramsay that there needs to be some form of disciplinary boundary in place to denote what is and what is not considered Digital Humanities, just as every other academic field has. This thesis accepts that the skill to code is a hallmark of the digital humanities, as it is a necessity for the well-rounded criticism of tools. However, it also feels there is much to be gained from the 'thinkers' position in regards to the relationship between DH and other humanities disciplines. The vision of Rockwell as the Digital Humanities analogous to statistics, an ancillary science, seems valuable for the relation between DH and History. History, in particular, has a long history of maintaining close relations with various ancillary sciences (e.g. Numismatics, Palaeography, Diplomatics, and Sillography) where both fields reinforce each other. However, both using the

[88] Rockwell, p. 253.
[89] Mark Sample, 'The Digital Humanities Is Not about Building, It's about Sharing', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 255–57.
[90] Sample, p. 257.

narrow and broad definition for the Digital Humanities, this thesis is unequivocally a DH project.

Conclusion

This chapter has discussed the existing scholarship relevant to this thesis from the variety of fields it draws from. It has not been possible to be comprehensive and review all literature in existence – no thesis can do so. It has, perhaps more so than other theses, reiterated some of the basics of these three fields; as it seeks to speak to readers from such disparate fields, this is also appropriate. It has discussed how each of these three fields, history, computer science, and the digital humanities, relates to the journey this thesis is about to undertake.

The historiographic background has laid down some red lines that the thesis needs to keep within to produce results that can be accepted by historiographical standards of evidence. It needs to be fully transparent in its tool design, both in regards to its operation and in its design, so other scholars may verify its findings – to do otherwise would be to create a 'black box'. A similar requirement exists for the data that is used with the tool: digital source criticism is key to the sound deployment of any tool this thesis produces. Meeting these two requirements allows the tool to be used as an extension of the historical method – in the understanding that while it can point towards historical truth, it can never be the sole evidence supporting it.

We have discussed the use of topic models and visualisations to investigate the historical questions informed by the historiography. Topic models have been used before by various historians, but their use has not been uncontroversial, as

they still suffer from issues related to transparency, in part as a result from their pedigree as search algorithms. This chapter has established that this thesis can make an original contribution to knowledge by providing a comprehensive and structured investigation into how to best employ topic models. In regards to visualisations, it has established that the key aspect of using visualisations is for them to expose aspects of the underlying data that would otherwise remain hidden; the data has to tell a story. It also found that heatmap visualisations have seen little use in the historical community, with existing work relying mainly on network visualisation. There is the opportunity for this thesis to make a significant and original contribution to knowledge by developing a tool that visualises the spatial nature of newspaper articles.

This thesis has defended its classification as a DH project. It established the field has no clearly defined and commonly accepted disciplinary boundary. However, even according to the narrowest definition of the Digital Humanities, that of the 'makers', this thesis still meets that standard for inclusion. It builds its own tools that are tailored to the data it intends to use, which allows for it to produce transparent scholarship and allow tool criticism. It will, in the next chapter, also engage in a critical review of those sources. After all, historical collections have a longer history than the day they were digitised, and this past needs to be taken into account before the tool takes shape.

Chapter 2: Archival Considerations

This chapter will discuss the material and technical aspects of the techniques this thesis has developed. Just as the sculptor's artistic vision is constrained by the shape and size of the marble, so too is the researcher limited by the data as to what knowledge they can generate from it. In other words, if something is not in the original data, it will not be reflected in the final product; historians are very familiar with this principle, as they often work with a fragmented historical record. The dataset underpinning this thesis is the *British Library Newspapers, 1800-1900* database, which contains 66 newspapers published in Britain during the nineteenth century. The dataset was released in 2003, but the story of its creation begins much earlier, more than two centuries ago.

This chapter will argue that understanding the processes that created and shaped the archive are critical for understanding the dataset. First, it will discuss the production of the data, including context of the press that created the newspapers in the archive, and how the practices in the industry at the time create peculiarities in the data, such as duplicate or highly similar texts. Then it will examine the retention of the finished newspapers in the archive, and how the ways in which the archive was handled created gaps in the archival record. Finally, it will discuss the two transformations the archive underwent. The first of these took place when the archive was transformed from a physical object – paper or microfilm – into digital data. The second was when this data was transformed and placed into the research setup used by this project. Each of these steps carry

implications for the form and quality of the data, and thus influence the possible uses of the archive. In the process of discussing these transformations, it will also touch upon two technological and methodological processes, Optical Character Recognition which was used to digitise the archive, and Keyword Searching which was used to create subsets from it for analysis.

During these discussions, this chapter will argue that constructing a tool and critiquing it needs to go hand in hand with knowledge about and critique of the archive from which it derives its data. We cannot use a dataset of Victorian periodicals digitised in 2000 without understanding its history before 2000 – it did not appear out of nothing. There was a physical archive before the digital, and it shaped the form of its successor. The entire pre-digital history of the sources we use in a digital form matters. Hence, it makes sense to discuss the creation of the newspaper articles themselves; the way they were stored; and the way they were digitised.

Construction

The story of the data in this archive begins when its principal components were first created: as an article in a newspaper, a century and a half ago. Between 1850 and 1900, the Victorian press exploded in a flurry of activity, as the abolition of 'Taxes on Knowledge' such as stamp duty made newspapers available to a larger number of people than before.[1] The number of newspapers sold has been estimated as growing rapidly, doubling in the fifteen years leading up to 1850 to

[1] Martin Hewitt, *The Dawn of the Cheap Press in Victorian Britain: The End of the 'Taxes on Knowledge', 1849-1869* (London and New York: Bloomsbury, 2014), pp. 165–66.

sixty-seven million, and continuing that growth for another twenty years.[2] One of the additional factors enabling this meteoric rise in the number of papers sold were the technological innovations that allowed for cheaper printing on cheaper paper.[3] The printing on cheaper, wood-pulp paper has significant implications for the conservation of the archive. This paper is more acidic in nature, which causes it to dissolve over time when in the presence of moisture.

For these reasons the number of titles in print, each title's circulation, and the diversity of content kept increasing throughout the fifty-year period under investigation. However, in the main, there are two kinds of newspapers that make up the archive. The first is the weekly paper, which was only published once a week (typically on Saturday or Sunday) and typically aimed itself at a primarily working-class audience. Such an audience only had Sunday off from work, and often had limited funds available to spend on luxury goods like newspapers, thus necessitating them being put in the market at a low price point.[4] For those that could not afford them even then, many towns had reading rooms geared towards the working class.[5] These papers aimed to give an overview and summary of the news of the week, as well as provide entertainment to the reader in the form of "populist and punchy human-interest stories".[6]

---

[2] Circulation is notoriously difficult to establish, see for estimates: Francis Williams, *Dangerous Estate: The Anatomy of Newspapers* (Longmans, Green, 1957), pp. 99–107; 103; Kevin Williams, *Read All about It! A History of the British Newspaper* (London and New York: Routledge, 2010), pp. 5; 54; 99; 110; Joel H. Wiener, 'The Nineteenth Century and the Emergence of a Mass Circulation Press', in *The Routledge Companion to British Media History*, ed. by Martin Conboy and John Steel (Routledge, 2014), pp. 206–14 (pp. 208–11).

[3] Francis Williams, pp. 121–24.

[4] Mick Temple, *The British Press* (Maidenhead: McGraw-Hill Education (UK), 2008), p. 23; Wiener, p. 207.

[5] Andrew Hobbs, 'The Reading World of a Provincial Town: Preston, Lancashire 1855-1900', in *The History of Reading, Volume 2: Evidence from the British Isles, c.1750-1950*, ed. by K. Halsey and W. Owens (Basingstoke: Palgrave Macmillan, 2011), pp. 121–38 (pp. 133–34).

[6] Temple, p. 26.

On the other hand, daily papers were typically addressed to a more middle-class audience, and focussed more on political and economic news. However, by design these were disposable items, intended to exist only for however long its news remained current. That said, while they were to an extent designed with obsolescence in mind, as the proprietor would wish to sell their latest issues, there is significant evidence that newspapers were often passed around, and retained (some form of) relevance for much longer.[7] Both types of paper in this archive were printed on relatively cheap paper; thinner and more fragile than the thicker paper used for books and more substantial periodicals. The archive is thus filled with newspapers that were produced at a low cost in order to compensate for a low cover price, produced at high volume on cheap acidic paper, and by a creator with a vested interest to sell the latest copy over readers retaining old ones. These attributes of a physical nineteenth-century newspaper make it fragile, even under the best archival conditions.[8] As we shall see later, in the case of this archive storage was not ideal, which had an impact on the quality of the digitisation.

Intended to be printed cheaply, the articles in these newspapers were, in most cases, written cheaply as well. News articles in this period of history had three main sources. The first were reports from other, often foreign, newspapers that were copied either in whole or assembled from its main points. As these papers needed to physically make their way to the United Kingdom, the news they carried

[7] Leah Price, *How to Do Things with Books in Victorian Britain* (Princeton: Princeton University Press, 2012), pp. 185–88.
[8] Nancy E. Gwinn, 'The Fragility of Paper: Can Our Historical Record Be Saved?', *The Public Historian*, 13.3 (1991), 33–53 (pp. 33–34) <https://doi.org/10.2307/3378551>; For a discussion on the economic incentives, see: Charles G. Steffen, 'Newspapers for Free: The Economies of Newspaper Circulation in the Early Republic', *Journal of the Early Republic*, 23.3 (2003), 381–419 <https://doi.org/10.2307/3595045>.

was occasionally weeks old when it reached the British readership. This kind of "scissors and paste journalism" was a major source for news from abroad, especially for the first 15 years this project looks at.[9] Even after the advent of the electric telegraph, many papers preferred to rely on this as a source for their international columns, as transmitting a news report, even in shorthand, was prohibitively expensive.[10]

As a way to address this problem, and decrease costs, the second source of international news was founded: the news (or press) agencies.[11] The best known of these, Reuters, began in 1851, and together with its competitors it quickly set the standard for overseas reporting.[12] These agencies, which a newspaper could subscribe to, were able to negotiate bulk rates for telegrams, spreading the cost over all subscribing papers. Whenever news arrived from overseas, the agency

[9] Bob Nicholson, '"You Kick the Bucket; We Do the Rest!": Jokes and the Culture of Reprinting in the Transatlantic Press', *Journal of Victorian Culture*, 17.3 (2012), 273–86 <https://doi.org/10.1080/13555502.2012.702664>; David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon, 'Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers', in *2013 IEEE International Conference on Big Data*, 2013, pp. 86–94 <https://doi.org/10.1109/BigData.2013.6691675>; Catherine Feely, '"What Say You to Free Trade in Literature?" The Thief and the Politics of Piracy in the 1830s', *Journal of Victorian Culture*, 19.4 (2014), 497–506 <https://doi.org/10.1080/13555502.2014.967545>; Stephan Pigeon, 'Steal It, Change It, Print It: Transatlantic Scissors-and-Paste Journalism in the Ladies' Treasury, 1857–1895', *Journal of Victorian Culture*, 22.1 (2017), 24–39 <https://doi.org/10.1080/13555502.2016.1249393>; M. Beals, 'Scissors and Paste: The Georgian Reprints, 1800–1837', *Journal of Open Humanities Data*, 3.0 (2017), 1 <https://doi.org/10.5334/johd.8>; Special Issue on the subject: Will Slauter, 'Introduction: Copying and Copyright, Publishing Practice and the Law', *Victorian Periodicals Review*, 51.4 (2018), 583–96; containing: M. H. Beals, 'Close Readings of Big Data: Triangulating Patterns of Textual Reappearance and Attribution in the Caledonian Mercury, 1820-1840', *Victorian Periodicals Review*, 51.4 (2018), 616–39.

[10] Tom Standage, *The Victorian Internet*, 3rd edn (London and New York: Bloomsbury, 2014), p. 150.

[11] James Mussell and Matthew Taunton, 'News Agencies', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 450–51.

[12] Donald Read, *The Power of News: The History of Reuters, 1849-1989* (Oxford, New York: Oxford University Press, 1992), p. (chap. 2); Scott Eldridge II, 'Change and Continuity: Historicizing the Emergence of Online Media', in *The Routledge Companion to British Media History*, ed. by Martin Conboy and John Steel (Routledge, 2014), pp. 528–38 (pp. 531–32); Jonathan Silberstein-Loeb, 'The Political Economy of Media', in *The Routledge Companion to British Media History*, ed. by Martin Conboy and John Steel (Routledge, 2014), pp. 75–84 (p. 81).

would deliver it to the papers at the local telegram rate.[13] As a result of this development, news could be brought to readers within 24 hours of it happening – providing, of course, that it happened near a telegraph station. Whether a newspaper received its copy by press agency or through scissors-and-paste, both have a similar effect on the composition of the dataset, which is centred around the article as its base unit: many articles are functionally identical, and this has implications for digital tools that aim to find similarities between different texts.

The third, and final source of news was the reporter or journalist, either employed directly by the paper, or more often an independent who submitted their copy hoping to be published. By necessity, these independent journalists reported on the more local issues, as they could not shoulder the financial risk of an overseas placement themselves. [14] These men, known as penny-a-liners for the reimbursement they received per line of copy supplied, provided a substantial amount of the content that made up a Victorian newspaper. Journalists that were directly employed were then free to be set to tasks directly supervised by the editor, such as an investigative project or an in-depth commentary on a significant issue.[15] There existed a definitive hierarchy amongst these journalists, and the latter were a rare and prestigious group.[16] This content is what a non-specialist would typically think of when they hear 'newspaper article', and is the kind of content that topic

[13] Roger Neil Barton, 'New Media: The Birth of Telegraphic News in Britain 1847–68', *Media History*, 16.4 (2010), 379–406 <https://doi.org/10.1080/13688804.2010.507475>; Wiener, pp. 211–12; Eldridge II, pp. 531–32.

[14] Wiener, p. 206; Matthew Farish, 'Modern Witnesses: Foreign Correspondents, Geopolitical Vision, and the First World War', *Transactions of the Institute of British Geographers*, 26.3 (2001), 273–87 (pp. 273–74) <https://doi.org/10.1111/1475-5661.00022>.

[15] Wiener, pp. 206–7; Randall S. Sumpter, '"Practical Reporting": Late Nineteenth-Century Journalistic Standards and Rule Breaking', *American Journalism*, 30.1 (2013), 44–64 <https://doi.org/10.1080/08821127.2013.767686>.

[16] Sumpter, pp. 44–45.

modelling is designed to work well on: unique documents with one or two central ideas, using relatively distinct vocabulary in their discussion of said ideas.

There was one other important type of content which is represented in the archive: advertising. Victorian newspapers were funded almost entirely through advertising income, which meant that an editor had to keep in mind the desire of his primary financiers as to placement in the paper.[17] In some cases, half the issue could consist of advertising; some of which ran on a recurring basis for weeks on end. The nature of the adverts changes substantially through the period under investigation. During the 1850s, they were mostly unique, placed by individual craftsmen and artisans; towards the end, adverts had entered the time of 'brand names' and large-scale and repeated signalling of the product's virtues.[18] Advertising was, as strange as it may seem today (where adblockers are used by a large number of internet users to avoid adverts online) not shunned out of hand by readers.[19] As a modern 'consumer culture' developed, people began to desire material goods – and adverts fed that desire.[20] Yet for all their popularity, adverts also highlight the ephemerality of newspapers, as they are often not retained in newspaper archives. Some titles published their adverts in special supplements, such as *The Times*, while other printed them on the brightly-coloured advertising covers used by popular weekly magazines. Examples of this include *Tit-Bits*,

[17] Francis Williams, pp. 50–51.
[18] Roy Church, 'Advertising Consumer Goods in Nineteenth-Century Britain: Reinterpretations', *The Economic History Review*, 53.4 (2000), 621–45 (pp. 631–34).
[19] Craig E. Wills and Doruk C. Uzunoglu, 'What Ad Blockers Are (and Are Not) Doing', in *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, 2016, pp. 72–77 <https://doi.org/10.1109/HotWeb.2016.21>; Church, p. 630.
[20] Lori Anne Loeb, *Consuming Angels: Advertising and Victorian Women* (Oxford: Oxford University Press, 1994), pp. 4–5.

Answers, and *Pearson's Weekly* in the final decades of the nineteenth century. Both these supplements and wrappers have rarely survived and are often excluded from digital collections.[21] But the form of adverts, as repeating texts, has important implications for the archive, as it means there will be many pieces of text that are identical to one another both across issues of the same paper, and across different papers, which poses problems for topic models.

Thus far, we have seen three kinds of ways in which an article may reappear in the archive as a different 'article' but with identical text: scissors-and-paste journalism, unembellished News Agency reports, and adverts. These can cause issues in the topic modelling phase, as they perform a statistical vanishing act when the model is created. As will be discussed later, topic models infer the importance of a word from how distinct it is. It assumes that the less a word occurs in a corpus, the more meaning it carries in the texts where it does occur, and thus the more important it is for classifying the text – and thus the more it is allowed to shape the model. Yet it also excludes words that appear either too little or too often across texts, as the meaning of these can't be inferred. In the former case because there are too little uses to draw from, in the latter case because they are so widespread as to not have a distinctive context of use. Thus, a situation where there are several copies of a text will either result in a single topic dedicated entirely to versions of that text, or text and its copies will be excluded entirely for being 'too common'.

[21] Francis Williams, pp. 81–82; Andrew King, 'Advertising', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 5–7.

For each edition, the editor of a paper had to decide how to arrange the building blocks of content: articles from correspondents and penny-a-liners, bulletins from the news agency and from foreign papers, miscellaneous literary extracts, letters from readers, special interest columns, advertisements, financial bulletins, and many more genres into a product that would sell to their prospective audience. The period that the sources in the archive originate from was the time of two very different systems by which these editors held their positions of authority. Some papers were run by 'sovereign editors', who were either both editor and proprietor of the title, or had a financial backer that was content to remain out of the way and leave them editorial independence. Editors of this type often had a strong political bent to their work, based in their own personal politics. One of the notable examples of this is W.T. Stead, who edited the *Pall Mall Gazette* between 1883 and 1889 with strong political fervour, responsible for one of the most well-known pieces of investigative journalism in the British Press, *The Maiden Tribute of Modern Babylon*.[22]

Other papers, instead of having the editor 'sovereign', were run by an editor working for a press baron; a single proprietor who owned a multitude of newspapers. These men often prescribed the political tone of their papers, and sought to make sure there was uniformity of political tone between their titles, often to further their own careers in Westminster. One of the men that illustrates this practice, although he reached the peak of his power only after the time period under investigation, would be A.C.W. Harmsworth, later known as Lord

[22] Francis Williams, pp. 126–28.

Northcliffe, who at the peak of his power was responsible for publishing not only *The Times* and *The Daily Mail*, but with his other holdings combined owned forty to forty-five percent of the daily newspaper market. Northcliffe was known for his anti-German stance and was a lead player in the 1909 Naval Scare that put Britain on a further collision course with Germany.[23] It was not uncommon for editors and proprietors to move on to or have a career on the side in politics. At the end of the period studied in this thesis, there were thirty proprietors and twenty-nine journalists serving as Members of Parliament.[24]

All of this means that the content in this archive is highly political, as either out of their own beliefs or under pressure from their proprietor, the editor made their content conform to certain political expectations. It also means that the material is politically distinct between papers, as the partisan position a paper took was part of its identity and would speak to a certain segment of readership at a specific time and in a specific place. This is extremely relevant to the case studies on imperialism, as several newspapers in this dataset had pronounced positions on imperial policy. *Reynold's Newspaper*, for example, was notably a radical paper home to opposition against imperial expansion.[25] Others, such as the *Daily Mail* and *Daily Telegraph*, actively advocated for imperialism.[26]

---

[23] J. Lee Thompson, 'Fleet Street Colossus: The Rise and Fall of Northcliffe, 1896-1922', *Parliamentary History*, 25.1 (2006), 115–38 (pp. 115; 119) <https://doi.org/10.1353/pah.2006.0011>.
[24] Kevin Williams, p. 110.
[25] Michael Shirley, ''Renold's Newspaper', ed. by Marysa Demoor and Laurel Brake, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 539–41.
[26] Nicholas Birns, 'Daily Telegraph', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 158–59; Kerry Chez, 'Daily Mail', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 157–58.

Retention

Having discussed the creation of the sources we study, and the way that practices in their creation may shape the archive and tools we build for said archive, we can now turn our attention to the way these documents have survived in the archive. This section tells the story of the way these newspapers were collected, stored, and in some cases, destroyed. All the events in the lifetime of these newspapers after the ink had dried, influence the way the archive looks, and what tools we build for it.

Once read, what happened to these papers? Most would be discarded, used as packing material or kindling.[27] But some were read and subsequently retained at the various libraries and reading rooms throughout the country. The latest issues of newspapers would more often than not be found in the reading room, while libraries kept bound volumes of back issues of the local paper as a source of local knowledge. While the 1850 Public Libraries Act allowed for the creation of libraries with public money, this was a slow process, and many cities did not have a central public library. For example, in the city of Newcastle-upon-Tyne there existed several subscription libraries, many specialised in a specific field of knowledge, alongside a large number of reading rooms which catered to more leisurely readers. It took thirty years for a public library to be established there, which subsumed all other major libraries in the city.[28] As a consequence, especially outside of London, newspapers would not normally have been retained beyond their use as local

[27] For a discussion of the afterlife of literature, see: Price, chap. 7.

[28] John C. Day, 'The Library Scene in an English City: Newcastle-on-Tyne Libraries, 1850-2000', in *Cambridge History of Libraries in Britain and Ireland* (Cambridge: Cambridge University Press, 2006), III, 206–15 (pp. 206–9).

sources, if at all. The retaining library would have had to collect the papers from the reading room, or take out a subscription themselves, to have the papers bound and take up valuable space.

However, the papers that make up the majority of the archive on which this thesis works came into the possession of the British Museum Library, the predecessor to the British Library, by a very different way. The British Museum Library itself is an anomaly, created out of the personal libraries of various English and British monarchs and supplemented with donations from a handful of notable antiquarians throughout the eighteenth century. Almost for a lack of a better place to put them, these incunabula and manuscripts were handed to the British Museum in Bloomsbury, which founded a Library department to hold and curate the collection. This library was chronically under-resourced, and much of its early funding in the nineteenth century came from the selling of 'duplicates'. While this did not affect the newspaper collections to a substantial degree, as these were not materials that would fetch much at auction, it does highlight the financially precarious situation in which decisions about conservation were made.[29]

By the middle of the nineteenth century, however, the British Museum Library was in a much better place, both financially through funding from the Treasury, and physically with the opening of a dedicated reading room and library in 1857. In line with subscription libraries, the British Museum Library viewed their newspaper collection as a record of the past that could be accessed to determine

[29] T.A. Birdell, 'The British Museum Duplicate Sales, 1769-1832, and Their Significance for the Early Collections', in *Libraries Within the Library: The Origins of the British Library's Printed Collections*, ed. by Giles Mandelbrote and Barry Taylor (London: British Library, 2009), pp. 244–46.

the 'true' events. In the words of John Winter Jones, the library's director between 1866 and 1873 and driving force behind a systematic acquisition policy for newspapers: "[I believe] that no source of information … is more important than the newspaper."[30]

However, for the first part of the time period this thesis covers, the acquisition of newspapers was not at all systematic, nor was it, in fact, directed by the British Museum Library. The institution did not actively pursue acquisition of newspapers apart from a number of London-based titles. Instead, it received the vast majority of its newspapers, most of them provincial, as donation from the Stamp Office, and later from its successor, the Inland Revenue Service. These institutions received a copy of each paper published directly from the printer, and retained it for two to three years in order to safeguard them as potential evidence in case libel proceedings were to be brought against the publisher.[31] After this time, they handed the material over to the British Museum Library. This mode of acquisition led to a minor crisis in 1869, when the abolition of stamp duty and a change in the copyright act no longer required newspaper publishers to deposit a copy with these services, and the British Museum Library did not have any funds earmarked to continue buying copies from the publishers.

The museum's solicitors agreed that the material could be claimed under the Copyright Act, which provided for legal deposit, but this would force the museum to accept all variant editions of national newspapers, as well as all the

[30] P. R. Harris, *A History of the British Museum Library*, 1753-1973 (London: British Library, 1998), p. 260.
[31] Harris, p. 271.

country papers which provided the same text under different titles. For example, the *Horsham Express*, *Chichester Express*, and *Worthing Express* were the same paper, but published under different titles in each of those towns. An alternative plan, to buy only those papers that would be of interest, was rejected, and between 1870 and 1872 the British Museum Library was hit by a deluge of material, which strained the capacity of the Newspaper Department to its limits. The problems were exacerbated by the library pre-emptively employing the London-based solicitor William Lethbridge to collect material from recalcitrant publishers, leading to a situation where there were two points of collection which did not fully communicate. This issue resolved itself when Lethbridge retired in 1873 due to ill health, but left a legacy of chaotic acquisition for these years.[32]

While the intake of papers stabilised in an organisational sense by the end of 1873, a new crisis was looming – one which would hound the British Museum Library, and particularly its newspaper section for the remainder of its existence. With an intake of twelve-hundred bound volumes of newspapers each year, occupying two hundred feet of shelving, space to physically store the papers was rapidly running out. An alarming note from the Director to the Board in 1874 noted that only eight more years of storage remained, at which point even the overflow storage in the old North Building Basement would be full.[33] However, as the British Museum Library's copies of newspapers were increasingly accepted as evidence in courts of law, the librarians noted that they could not simply get rid of

[32] Harris, pp. 271–72.
[33] Harris, pp. 272–73.

old material.[34] Therefore, a dedicated newspaper reading room would be built in the south-eastern corner of the site, which would become the White Wing Reading Room. This extension was completed in 1885, and in order to maintain the collection while construction was ongoing, only the morning edition of papers with more than one edition would be retained.[35] It would thus be dangerous to assume that the composition of the archive is an accurate reflection of the press landscape of the nineteenth century. A reader at the newsstand would have experienced a continuous cycle of editions that each added to their predecessor; we do not have that option, as the rest of the day's issues were never retained.

Yet again, space and cost were an issue. While there was now enough space to deal with the anticipated newspaper input for the foreseeable future at the then current rate of acquisition, and the amount of daily visitors to the newspaper reading room rose from thirty to fifty between 1886 and 1890, now it was becoming prohibitively expensive to bind all the newspapers.[36] One assistant librarian complained that the binding was twice as valuable as the papers it contained: a year's issues of a penny paper only amounted to twenty-six shillings, binding it in three volumes cost fifty-four.[37] Thus, the decision was taken to only bind the most used papers and the London dailies, while keeping the provincial papers as loose-leaf packets tied with twine. However, the binding of these titles was resumed in 1889, as readers had been complaining of damaged and missing

[34] Harris, p. 344.
[35] P.R. Harris, 'The British Museum Library, 1857-1973', in *Cambridge History of Libraries in Britain and Ireland* (Cambridge: Cambridge University Press, 2006), III, 281–98 (p. 283).
[36] Harris, p. 363.
[37] Harris, p. 276.

issues in the bundles they retrieved from the reading room staff.[38] The issues that were lost in this way were never systematically replaced. This contributes to gaps in the archive prior to that date, as reported by Nicholson.[39]

The observant reader may begin to note a pattern in the history of the newspaper collection as by the turn of the century the White Wing, which had been anticipated as providing newspaper storage until the 1920s, was running out of space due to more titles being acquired than planned for.[40] In 1897, the binders were moved out of the North Building Basement so it could, once more, be used as newspaper store, but all parties involved knew that the problem would recur, and drastic action was sought. In the interim, the storage in the North Wing Basement caused the newspapers to deteriorate due to moisture and pollutants in the air. This process of acidification poses a problem for all wood pulp-based untreated paper, but the presence of moisture and pollutants speed up this process.[41]

As the provincial papers were the least consulted by a wide margin, the Principal Librarian suggested that these be transferred over to local institutions, such as the newly formed county councils. This would free up over eight hundred shelf presses on the outside of the round reading room for London papers. However, in order to allow for this transfer of material the Copyright Act would

[38] Harris, pp. 344–45; 348.
[39] Bob Nicholson, 'Looming Large: America and the Late-Victorian Press, 1862-1902' (unpublished PhD Dissertation, University of Manchester, 2012).
[40] Harris, III, p. 285.
[41] H. Cheradame, S. Ipert, and E. Rousset, 'Mass Deacidification of Paper and Books. I: Study of the Limitations of the Gas Phase Processes', *Restaurator*, 24.4 (2008), 227–239 (p. 227) <https://doi.org/10.1515/REST.2003.227>.

need to be amended. Legislation to this purpose was drawn up between 1898 and 1899, but with the Salisbury Government occupied with foreign policy, and in the face of opposition from the Commons, this legislation was quietly withdrawn in June 1899. This left the British Museum Library to search for new avenues to store its growing newspaper collection, but the failure of the Bill in Parliament did make the Treasury more susceptible to requests for funding, as their refusal to pay for any new construction until the 'question of disposing of newspapers was fully considered' was a driving force behind the legislation.[42] Thus, funding was made available for the building of a dedicated newspaper storage depot at Colindale, in the borough of Hendon on the outskirts of London. Construction on the five and a half acre site finished in 1905, and a massive operation to transport the bound newspapers by van began soon after. It transported nearly the entire collection of provincial papers, totalling 53,465 volumes to the new site, from which readers could request them to be brought to the Newspaper Reading Room in Bloomsbury. The fact that the collection did stay together in the composition it has now, instead of being spread out across the counties to county councils, who may have taken completely different decisions about their conservation over the past century, is thus entirely down to the political landscape of the 1900s.

Collindale remained the main site for newspaper storage throughout the first quarter of the twentieth century, but the overrunning cost of building works on the King Edward Galleries at the British Museum made the Treasury reluctant to fund further expansion on the site when the repository filled up in 1913, instead

[42] Harris, pp. 376–77.

forcing the collection to be divided between Colindale and the basement of the new galleries.[43] Requests for funding once the building works in Bloomsbury were finished were turned down due to the First World War, and it wasn't until 1926 that funding was secured to build a second building with dedicated newspaper reading room at the Colindale site.[44] While the British Museum Library had weathered the First World War without suffering any physical damage to its holdings, it was less fortunate in the Second World War, with the Newspaper Collection in Colindale being hit particularly badly. A major German raid targeting the nearby Royal Air Force Station at Hendon Aerodrome and industrial estates in the area on the 23rd October 1940 struck the newspaper storage with a mixture of high-explosive and incendiary ordinance. Several thousand volumes of bound Irish and Provincial Newspapers burned to ash, while up to forty-thousand volumes were recorded as having suffered some kind of scorching or water damage.[45] This totals up to nearly the entire Colindale holdings being affected in some way by the Luftwaffe. In other words, our analysis of the data derived from the archive now is shaped by the aim – or lack thereof – of German bombers in the Blitz.

The Colindale repository was rebuilt in 1957, but many of the volumes were found to have been suffering from deterioration already before the bombing, probably as a result of the storage conditions in the North Wing Basement. The age of the low-quality paper and cheap ink made pages illegible and fragile. In response to these findings, from 1963 onwards the British Museum Library began

[43] Harris, p. 378.
[44] Harris, III, pp. 288–89.
[45] Harris, III, p. 292.

to convert parts of its stock to microfilm, beginning with those issues that had been worst affected by the war, and the 1913-1927 collection of provincial papers that had been at the Edward VII Galleries basement so it could be made accessible from Colindale. This filming programme has had a knock-on effect on the digitisation, which happened from microfilm. As film is digitised much quicker than paper, the availability of these films gave the titles for which film was available priority for scanning.

After these war-damaged volumes, priority was given to high-demand or at-risk items, with the aim being to convert the entire collection. However, as this same microfilm department was also responsible for microfilming items for the Manuscript and Rare Books Department and the General Library Department, as well as producing microfilm copies of modern papers under the NEWSPLAN programme, this aim was never quite achieved.[46] The long time-range over which this microfilming took place means that there is significant variance in the quality of the microfilm itself between issues and between papers within the archive, even if the difference in quality of the original material is taken out of the equation. Thusly, the material artefacts underwent their first transformation, from paper to picture. However, some of the limitations of the original papers remained in place: only one person could view one paper at a time; pages had to be read sequentially (or at least skipped in sequence); and the entire page had to be read to avoid missing information. In response to these limitations, the British Library (as the successor

[46] Harris, III, pp. 294–96; John Hopson, 'The British Library and Its Antecedents', in *Cambridge History of Libraries in Britain and Ireland* (Cambridge: Cambridge University Press, 2006), III, 299–316 (p. 307).

institution to the British Museum Library after its separation in 1972) began a programme seeking to digitise its newspaper holdings.[47]

Transformation

The story of the newspaper articles has now spanned almost 150 years. We have seen the article being created in the context of the various systems of praxis in Victorian journalism, and have observed its battles for survival against the forces of time (and a lack of funding) as a part of the British Museum Library and British Library. We now come to the final part in the story of the archive so far: its transformation from a physical object to digital data. For this discussion there are four items that need to be discussed. First, this project has to discuss the process of Optical Character Recognition (OCR), by which the scans of the pages or microfilm are transformed into machine-readable text. This is intended to show the state of the OCR technology at the time the archive was created, and discuss the way in which it influences the digital version of the archive. The second part will cover the specific case of this archive's digitisation, including what is known about choices made during the selection of the titles included, up to the point where the data was handed over to this research project. Third, there will be a discussion of the way this project transformed the data and set up its infrastructure for creating subsets to use in visualisation and topic modelling. Finally, the fourth section will discuss the use of keyword searches as a tool for generating subsets of the data and evaluate their methodological implications.

[47] Edmund King, 'Digitisation of Newspapers at the British Library', *The Serials Librarian*, 49.1–2 (2005), 165–81 (p. 167) <https://doi.org/10.1300/J123v49n01_07>.

OCR: Reading without understanding

At this point, it pays to make a short diversion into the realms of text digitisation and optical character recognition (OCR), in order to highlight its operation, its defects, and why these known issues were considered acceptable by the archivists in charge of digitisation. Optical character recognition came about as the solution to the problem that there was a large amount of text which was not digital, but which had to be made available to computers. Storing text in a digital format and using this text as an index has been done since the 1950s. The pioneer in this field was Fr. Roberto Busa's *Index Thomesticus*, which contained the entire text of the works of Thomas Aquinas. However, the desire to provide a perfect copy, which could be used for computational linguistic analysis meant it had to be rekeyed by hand from the physical sources, a process which lasted thirty years – it started on punch card and ended on CD-ROM.[48] Such a timescale was unacceptable for digitising the much larger body of texts that were being processed during the digital boom of the 1990s. For them, optical character recognition software was adapted from its original purpose – reading passports and ID cards – and used to rapidly digitise large bodies of text.

Optical character recognition operates by detecting the smallest unit of text, the character, on the image it is given. It then compares this small area, only a few dozen pixels across, with its database of character shapes, before determining with a certain degree of certainty what character it is dealing with. It writes these two outputs, determined Unicode character and certainty value to an output file, and

[48] R. Busa, 'The Annals of Humanities Computing: The Index Thomisticus', *Computers and the Humanities*, 14.2 (1980), 83–90 <https://doi.org/10.1007/BF02403798>.

then repeats the same process using the next character on the image. This is a necessity as for a computer, a character in an image is nothing but a collection of pixels, while a character in a word processor is a binary value representing a character in a table of characters, and changing from one to the other is difficult. Thus, there are two factors that determine how good the output of the OCR software will be: the quality of the image, with a higher-resolution image giving better results, and the level of sophistication in the matching algorithm, of which a large variance exists that this thesis will not explore in detail.

Thus, the determining factor for OCR quality, from the perspective of this archive, is image quality. This encompasses a wide variety of factors, including but not limited to: the sharpness and focal depth of the image; the presence of damage on the physical object; ink bleeding through from the recto side of the page; the clarity of print on the original object; and if digitised from microfilm, the sharpness and focal depth of the microfilm camera, as well as any damage or deterioration of the microfilm itself. Any of these factors may adversely affect the quality of the OCRed text, and validating the results of the process is difficult, as such an undertaking would need a perfectly accurate digital version of the text to compare with the version generated by the OCR. However, if such a version did exist, there would be no reason for creating the computer-transcribed text. This means that any such research, by necessity, works by taking samples. The problem, however, remains. A representative sample, or baseline, on a corpus of ten or twenty million texts is still tens of thousands of documents that need to be transcribed with near-

perfect accuracy. The financial and time investment required for such research is simply too high to bear for most research projects.[49]

Alternative methods to full-text 'gold standard' comparison have been found. By comparing individual words with a dictionary, giving an estimate of the word-level accuracy; compared to full text comparison, which delivers character-level accuracy estimates. This method was used to investigate the error rates on *The Times Digital Archive*, and found error rates around 30% for a comparable time period as this thesis covers.[50] Research using this method on comparable Canadian newspapers showed similar average error rates, although it must be noted that the quality distribution over the dataset is non-Gaussian; that is to say, certain characters are much more likely to be transcribed incorrectly than others.[51] For example, on material using a roman font, like most Victorian newspapers, the character 'h' is often incorrectly transcribed as 'b'. Users of similar datasets produced by other national libraries and archives have also investigated the data quality of OCR-digitised newspaper archives, often as part of a preliminary study before large-scale textual analysis projects. In a 2016 review of the OCR quality of the Finish national newspaper archive, digitised around the same time as that of the British Library, researchers found error rates hover around 30%.[52] An earlier review of the British Library Newspaper collection also reported error rates

[49] Kimmo Kettunen and Tuula Pääkkönen, 'Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means', 2016.
[50] Kai Niklas, 'Unsupervised Post-Correction of Ocr Errors' (unpublished Master's thesis, Leibniz Universität Hannover, 2010).
[51] Beatrice Alex and John Burns, 'Estimating and Rating the Quality of Optically Character Recognised Text', in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, DATeCH '14* (New York, NY, USA: ACM, 2014), pp. 97–102 <https://doi.org/10.1145/2595188.2595214>.
[52] Kettunen and Pääkkönen, 'Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means'.

between 22 and 32%.[53] This corresponds to the state of the art in OCR-based digitisation at the time.[54]

However, the OCR quality in the data is far from uniform, and varies wildly from paper to paper. As shown by the graph below, there is a wide variety, with some papers suffering from 60% error rates. In general, character error rates hover about 30%, which is acceptable for human readers, who are able to reconstruct the original word without too much difficulty (figure 2.1). It is also no major problem for indexing scanned archival pages, as mistranscribed characters in the index can be covered by a fuzzy search function. Indeed, this is the solution that the British Library adopted in this project.[55] For topic models or any other computer-assisted discourse research, OCR errors pose a much more significant hurdle, as they rely on accurate representations of language. An example of the extremes of OCR quality found in this archive is attached as appendix 1.

[53] Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', D-Lib Magazine, 15.7/8 (2009) <https://doi.org/10.1045/july2009-munoz>.
[54] Edwin Klijn, 'The Current State-of-Art in Newspaper Digitization: A Market Perspective', *D-Lib Magazine*, 14.1/2 (2008) <https://doi.org/10.1045/january2008-klijn>.
[55] Edmund King, 'Digitisation of Newspapers at the British Library', pp. 175–76.

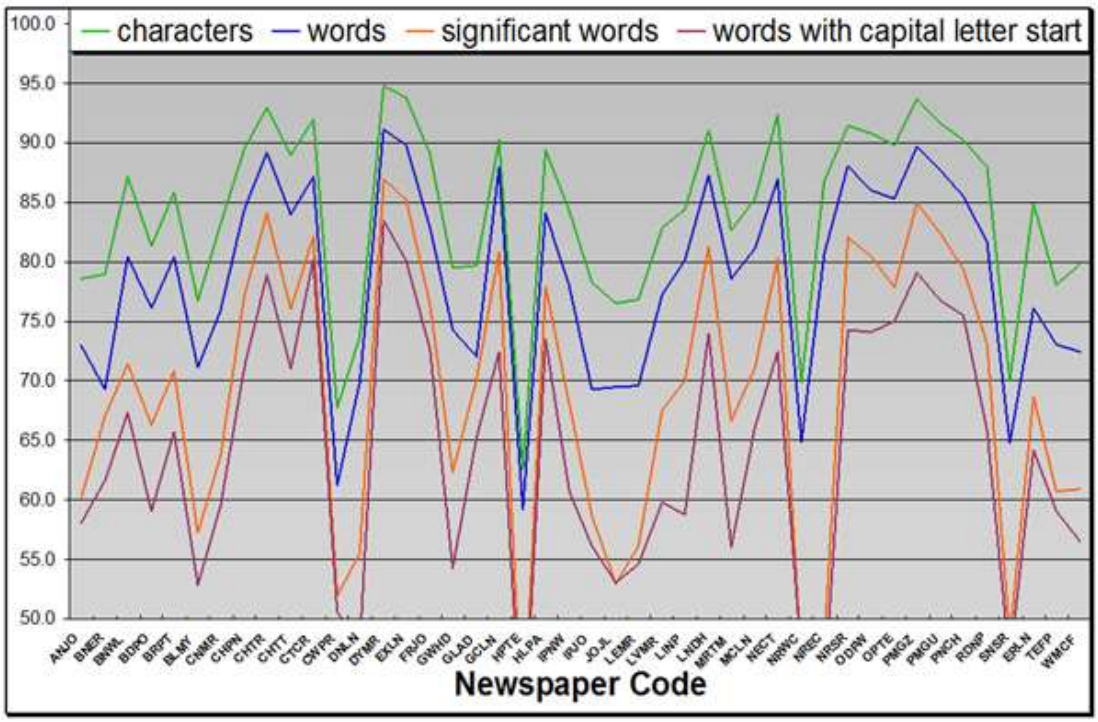


Figure 2.1: Error rates in the British Library Nineteenth-Century Newspapers, Part I.[56] *The variance in OCR quality is significant, between 65% at worst and 90% at best for individual characters.*

Although there had been experiments in the 1990s with the scanning and OCR using the British Library's *Burney Collection*, digitisation of the newspaper collection truly began in earnest in 2001 with the arrival of the *British Library Newspaper Pilot*. Partially a response to the commercial digitisation of *The Times* by Gale Cengage, this project was conceived as a testbed for the technologies that would be needed to process the incredible amount of newspapers held by the library, selecting for digitisation a few months for each year for a handful of titles. As a feasibility study, the pilot focussed on speed, and used only microfilmed papers that could be fed into the scanner as a reel. Anything that could conceivably be automated, was, including the page segmentation, which used "custom patented artificial intelligence fuzzy search technology".[57] Already in this project, issues with

[56] Tanner, Muñoz, and Ros, 'Measuring Mass Text Digitization Quality and Usefulness'.
[57] Edmund King, 'Digitisation of Newspapers at the British Library', pp. 173–74.

the quality of the OCR were noted, but as the text was only intended to act as an index to the page images, this was considered acceptable by the project leaders and the head of Newspaper Collections. It was also determined that the root cause for the lack of OCR quality was the poor quality of both the original material and microfilm surrogates, although other issues were caused by obsolete characters or odd fonts in early texts, such as the long 's' or gothic type.[58] The issues stemming from misread text were mitigated by implementing a fuzzy search function. It also included a separate XML layer to capture the placement of each element of text on an image, in order to direct the user to the place their chosen keyword occurred. Next was the *Collect Britain Project*, which was geared towards digitising the full run of one single paper, including its ephemera. This paper was the *Penny Illustrated Paper*, chosen for its wide range of layouts and content. Digitising this title led to further refinement of the processes and was complete by 2003.[59]

Physical to Digital

The digitisation project that contributed most of the data this thesis draws on, however, was the JISC-funded *British Newspapers 1800-1900* project. In total, the operation digitised over two million pages of newspapers, though this number covers just forty-seven titles: a fraction of the British Library's entire holding. Where possible, the digitisation effort was carried out on new microfilms of the physical papers, in order to provide consistent images – in only ten percent of the cases old film was used due to the original papers having since been destroyed completely or having deteriorated to the point that no better images could be

[58] Edmund King, 'Digitisation of Newspapers at the British Library', p. 173.
[59] Edmund King, 'Digitisation of Newspapers at the British Library', pp. 172–76.

taken.[60] This new digital archive was aimed particularly at higher- and further education providers, and was the first digitisation project undertaken by the library where such institutions were part of the design process.[61] As a result, the focus was mainly on creating a widely-ranging collection that would fit as many different curriculum requirements as possible. In a counterpoint to this decision, some special collections were created within the broader dataset. For example several Chartist newspapers were specifically selected because they are a-typical and would prove useful for nineteenth-century history courses.[62] When the initial phase of digitisation was wrapping up in 2006, the British Library and JISC agreed for additional funds to be made available to digitise an additional million pages, consisting of nineteen mainly provincial titles that had been selected for inclusion in the initial phase, but were dropped to allow for more 'national' papers to be included. Together, these two phases of digitisation created the *British Library Newspapers, 1800-1900: part I and II*, for a total of 69 titles. Both instalments are contained within the dataset used by this project.

The requirements of the project for teaching and learning needs necessitated it to be a fully online product, in order to make it accessible to lecturers and students around the country. Initially, the aim was for the entire collection to be accessible freely online via institutional access – a condition stipulated by the

[60] Edmund King, 'Digitisation of British Library Newspapers 1800-1900', *British Library Newspapers*, 2007 <http://find.galegroup.com/bncn/bncn_01.htm> [accessed 18 August 2019]; Edmund King, 'Digitisation of Newspapers at the British Library', pp. 179–80.
[61] Edmund King, 'Digitisation of Newspapers at the British Library', p. 178.
[62] 'British Library 19th Century Newspapers : JISC', 2008 <https://web.archive.org/web/20080918080147/http://www.jisc.ac.uk/whatwedo/programmes/digitisation/bln> [accessed 18 August 2019].

principal funder JISC.[63] The commercial exploitation of the data would be handled by Gale-Cengage, who partnered with the library because their experience with the *Times Digital Archive* had shown they were well-equipped to handle this process. [64] The digitisation was undertaken in two phases, which results in two image qualities in the archive. After digitisation the articles and their metadata were stored in the proprietary GIFT format.[65]

Three years after the end of the collaboration with the American company, the British Library also sold the use of the data to the Brightsolid group, later known as findmypast.com, in a joint project to further digitise its newspaper collections and bring them outside of academic use. In contrast to the sale to Gale, the collaboration with findmypast led to a project that the British Library retains a stake in, namely the British Newspaper Archive.[66] This timeline is somewhat muddy though, as few of the original press releases that could have clarified who partnered with who at what time and under what conditions have survived online.

However, the important aspect for the dataset used in this thesis is that the condition set by JISC that the data should be made accessible to researchers is still valid. The *British Library Nineteenth-Century Newspapers* dataset is still accessible

---

[63] 'British Library 19th-Century Newspapers : JISC', 2010 <https://web.archive.org/web/20100607033152/http://www.jisc.ac.uk:80/whatwedo/programmes/digitisation/bln> [accessed 18 August 2019].

[64] '300 Years of the British Press Goes Digital – Gale and the British Library Build a Digital Reading Room', Gale-Cengage Press Releases, 2008 <https://news.cengage.com/library-research/300-years-of-the-british-press-goes-digital-%e2%80%93-gale-and-the-british-library-build-a-digital-reading-room/> [accessed 18 August 2019].

[65] See for recent work on this topic, and for comparison with other digital archives: M. H. Beals and Emily Bell, 'British Library 19th Century Newspapers', in *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges* (Loughborough, 2020) <10.6084/m9.figshare.11560059.>.

[66] 'British Library and Brightsolid Partnership to Digitise up to 40 Million Pages of Historic Newspapers', British Newspaper Archive Press and Media Information, 2011 <https://web.archive.org/web/20110526213116/http://www.britishnewspaperarchive.co.uk/archive-media.php> [accessed 18 August 2019].

royalty-free for every UK HE institution, though Gale-Cengage charges for the handling and shipping of the drives. Thus, this thesis was able to contact Gale-Cengage and, after some discussion, purchase the entirety of the *British Library Nineteenth-Century Newspapers: part I* and *II* in its raw XML and TIFF format which was subsequently used to build the datasets used in this thesis.[67]

Research Server Setup and Retrieval

When our dataset arrived, it was on four two-terabyte hard drives. These contained, for each page of the newspaper archive, two files. The .tiff images of each page scan, and an .xml file with the automatically generated transcription. The first of these was not used for the topic modelling, as it contains no textual data, but the second one needs to be transformed in order to be usable. The xml file structure subdivides a page into columns, articles, lines, and words; each with their coordinates on the image of the page given.[68] This makes it useful for the purposes of the archive curators – as an index – but it needs to be re-shaped to be used for topic modelling and historical research. At the least, this re-shaping had to overcome the most significant problem with the xml format: that articles that spanned over several pages were not considered as one document. For this purpose, the xml files were converted to JavaScript Object Notation (JSON), which allowed them to be indexed on an Elasticsearch Server. This server setup consisted of a WD PR4100 four-bay NAS[69], with each of the drives holding one 2TB drive for storing the JSON version of the archive. In case of drive failure, the drives with

[67] Beals and Bell.

[68] Beals and Bell.

[69] Network-Assisted Storage. A NAS is a step up from an external Hard Drive, as it contains its own computer solely for managing access requests to the multiple drives and facilitating data transfer. The number of bays defines the number of 3.5 inch HDD's that can be mounted.

the xml files were retained separately, so they could be used to reconstruct data lost should a drive in the NAS break.

The server setup initially used a HP Z220 with an i5-2500 processor and 4 GB DDR3 RAM as a host for access and querying. This machine was used because it was the standard desktop computer supplied by my university at this time, and I hoped to demonstrate that relatively advanced computational analysis could be performed using hardware sitting on the desk of a typical humanities researcher, but the performance of this machine was severely lacking. A query for all the articles containing a single keyword in a year could take up to two days to run, which limited the scope of this project and disincentivised the use of speculative queries in favour of searches that were more likely to generate useable results.

Later on in the project, this machine was upgraded to a HP Z420 with an E5-2680 processor and 64 GB DDR3 RAM, which provided a major boost in the speed at which the data could be queried, and therefore enabled me to run more queries and pursue new research questions. Crucially, this machine was not prohibitively expensive, demonstrating that it *is* possible for humanities researchers to undertake this kind of research using higher-end desktop-class machines, rather than fighting for access to their university's oversubscribed High Performance Computing (HPC) services. The arrival of this machine meant that queries could be done on more speculative keywords, as rather than taking days and potentially finding nothing of use to the case study, we now had a way to run queries in hours, where the risk of finding nothing balanced against the time invested. When

designing a research project using data on this scale, the time constraints, as dictated by the hardware used, shape the questions it is feasible to answer.

Having covered the hardware above, we can now briefly mention the data setup. The database uses three different indices, each on a different aspect of the source material, representing a different level of 'zoom'. These are the Article, Image, and Issue indices, which can also be found in figure 2.2 These are independently searchable, but the way the Id's are interconnected means going from a specific to a more general level is easy, while narrowing down from an issue to an article is possible by relying on the folder in which the pages are stored. As part of the indices, lookup and reverse-lookup was implemented to allow flexibility of use. The choice of these indexes itself does influence the questions the project can ask of the data: for example, as the 'Article' index does not include a 'date' field, we can't ask for all article texts for a given date directly – we have to use the Id to bridge that gap.

Index	Field	Contents
Article	coords	Coordinates of the article on the page
	fulltext	Title + text generated by OCR
	id	Identifier string, made up of: Archive part + abbreviation + date + issue + page number + article number
	issue_id	Identifier string of the issue this article is in
	newspaper	Name of the paper as on the masthead
	pages	Pages this article covers
	src	File location of the xml file containing the unprocessed OCR data
	text	OCR-generated article text
	title	Human-corrected article title
	type	Type of article
Image	id	Identifier string, made up of: Archive part + abbreviation + date + issue + page number
	src	File location of the page scan of this page
Issue	abbreviation	4-letter code representing the paper's title, consistent over name changes
	date	Date of issue
	day_of_week	Day of the week of that particular date
	id	Identifier string, made up of: Archive part + abbreviation + date + issue
	issue	Integer representing morning and evening editions on a same day
	language	Language of the text
	name	Name of the paper as on the masthead
	pages	Number of pages in the issue
	source	What medium the issue was scanned from
	sourceId	Alphanumeric code reparenting the part of the archive the issue was digitised for
	src	File location of the folder containing the pages in this issue
	volume	Which bound volume the physical paper is part of

Figure 2.2 Format of the information retained from an article in the indices of the database. Each id contains the information of the indices below it, allowing an 'Article' to be linked to a 'Page' and 'Issue'.

The way this thesis transformed the dataset have thus been minor but significant. It did not add or remove information, and thus has no influence on the composition of the archive. However, it did transform the shape of the data significantly; from an xml-based index designed to highlight individual words to a

json-based record containing its full text. By placing the articles in the indices chosen, this project made a choice about the way it wishes to interrogate the archive; in this form, it gives privilege to the article as the building block of the paper.

Keyword Searches for Subsetting

Having explored the transformation this project imposed on the archive, and the way it chose to structure the dataset for its own research purposes, all that is left is to set out the way the corpus is subset for topic modelling and visualisation. This is the process by which the archive, which acts as the corpus for our purposes, is subset into smaller collections. The term subset is appropriate over subcorpus, as the way these articles are selected is not intended to create a selection that is representative.

The subsetting is necessary because the total collection archived articles is still too numerous to feed into a topic modelling program at this point; over 23.3 million articles. Topic modelling this amount of text would not only take an unreasonably long time, there would also be no guarantee that a topic corresponding to a historical research question would form. Therefore, a keyword search was used to create a smaller subset of texts on which to run the topic model, specific to the case studies that this thesis undertook. Discussion and justification of the choice of search terms, therefore, will be done in the relevant case studies. This would expose the topics present in the text that mentioned the keyword or set of keywords, allowing the context in which keyword(s) were used to be

explored. The choice to employ keywords caries some methodological implications.

Especially in the humanities, relying on keyword searches has been criticised. Ted Underwood likens searching for sources based on the appearance of keywords to a "Boolean fishing expedition for sources that may or may not exist", continuing to observe that "[keyword]search is not just a finding aid; it's analogous to experiment—although, to be sure, there's something a bit dubious about experiments that get repeated until they produce a desired result. The search terms I have chosen encode a tacit hypothesis about the [topic], and I feel my hypothesis is confirmed when I get enough hits."[70] Essentially, most users of keyword search adapt their search terms until they have the desired amount of sources for their close reading step: depending on text length generally between ten and fifty. Any less and there may not be enough examples to build a conclusion from, but when keywords return too many hits, the original problem that the keywords were deployed to defeat – too many sources – still exists. This results in a high risk of confirmation bias: after all, in a database of tens of thousands of sources, finding twenty examples of a keyword used in a certain way is trivial.

Proving the representativity of the selected sources is the problem when conducting a close reading, as doing so with a large amount of sources defeats the purpose of keyword-based selection. In a big data approach such as the one used in this thesis this is expected to be much less of an issue; while we still aim to reduce the size of the dataset before us, we merely reduce it to a scale that our

[70] Underwood, pp. 64–65.

computer can handle, which is orders of magnitude more than a human researcher. As the results from the keyword search were to be fed into the topic modelling program, this thesis could allow itself to keep its search terms broad; the specificity sought is put in place by the topic models. In general, terms (or collections of terms) were chosen that led to subsets of at least 10,000 articles per decade, which provided plenty of material for the topic models to be generated from. These are also big enough that there is little to no risk of 'cherry picking' results out of a 'boolean fishing expedition'. However, it does mean that the topic models this project uses are more directed than those other researchers have used in the past, yet as it is simply not feasible to model the entire corpus and hope relevant topics come out, this project believes this to be an acceptable solution.

The actual keyword searches were handled by an ElasticSearch host running on the Z220/420, queried over WIFI through a Python script on a Macbook Air, which in turn had a removable external hard drive connected to store the JSON files returned by the search server. This script took as input the keyword(s), in addition to any other selectors, such as the amount of articles required or if they need to come from a particular newspaper. It then formulated the queries in Elastic, while dividing the total amount requested in smaller batches to ease the load on the server and lessen the impact of a dropped connection. These smaller, 100-article responses were organised by the script into a JSON file containing all the responses. For each article, this file contained the title of the newspaper, date of publication, the manually-corrected article title, and the full text of the article generated by OCR. This process is shown visually in a flowchart in Figure 2.3.

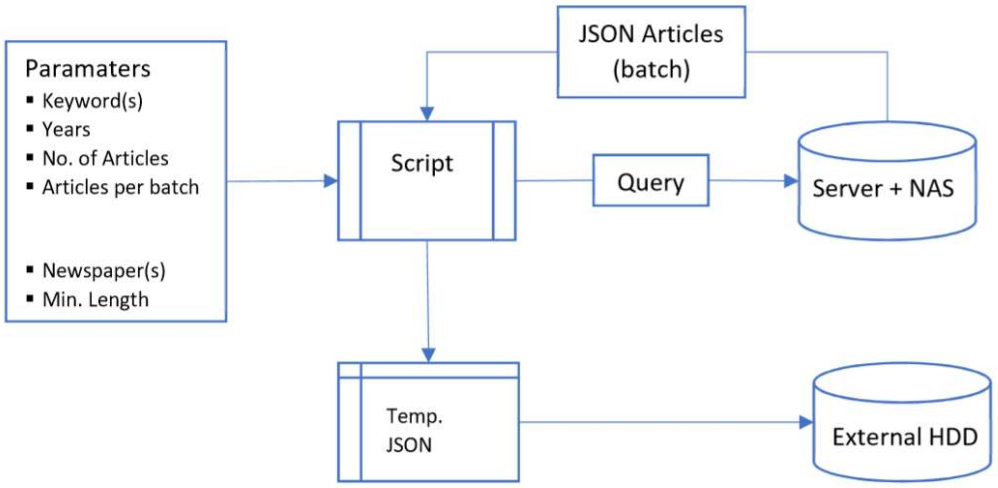


Figure 2.3: A flowchart representation of article retrieval. The querying script is fed a number of parameters. These are parsed into an Elasticsearch query, which is sent in batches to retrieve the articles from the Server-NAS combination. It temporarily stores the articles, before collating the batch responses into one JSON file, which is written to an external HDD.

The file name of the collated files reflects the information about their contents that was most important to this project: the keyword which had generated the subset and the year that the articles found were from. Optionally, there was room for an extra modifier such as a specific newspaper title. These files could then be loaded into the topic modelling tool. If intended to be visualised, they could be used as inputs for a second search query, which retrieved the additional information the visualisation process needed.

Conclusion

This chapter has shown that the material that this thesis uses did not appear out of thin air, but instead has a rich and storied history of its own. It was created by certain people at a certain time with a specific purpose, using particular methods, which led to a product that has a particular form geared at lowering costs of production, and with repetitions in its content. It was subsequently preserved, initially more by happenstance than through any predetermined strategy, by an

institution that was not always sure if it wanted these sources in its collection as they hogged valuable space and money. There are missing pages from the days the local papers were kept as bundles, and missing volumes due to the Luftwaffe raid on Hendon. The newspaper's material, which was never high-quality to begin with, was not improved by being kept in damp basements at various points in its life. This led to microfilming in an effort to preserve the material, but these microfilms themselves are not always of the highest quality due to their production over a timeframe of a quarter-century by an overworked department. Once digitised, the papers were indexed based on imperfect character recognition software, but these flaws were accepted as they produced a text that was adequate for what it needed to do. They were then transformed by the research project to fit its goals, and subsets of it were created by keyword searches which, due to the OCR, may or may not find all mentions of their keyword. Yet it shall have to do: the archive is imperfect, but it is the best it is going to get. Understanding the imperfections of both the archive and the processes involved in transforming it, however, is crucial for using it as a source for research – historical or otherwise.

Chapter 3: Topic Modelling

Having discussed the material on which we work, we can now start chipping away at its rough and unhewn surface to create a piece of art. But with what? What tools are appropriate for this particular material, and the particularities of historical research questions? The first of the chisels that this project will use is Latent Dirichlet Allocation (LDA), a form of topic modelling. Understanding the operation of LDA is crucial to using it as a research tool; this includes an understanding of the way in which the probabilities underlying the model are calculated. As this project produced its own tool to apply LDA , it will have to discuss the way this tool works. This means critically discussing four steps taken by the tool: the way data is pre-processed before LDA is applied; the way LDA modelling is influenced by parameters used; the way the model can then be returned to a human-readable form and viewed; and the way these models are then analysed and have meaning ascribed to them.

A large part of the work in this chapter is methodological. Throughout the discussion of the capabilities and operation of the tool it will continuously argue that, if the object of developing this tool is historical study, it has to shadow historical methodology. This means that from the start, it has to accept the absolute binary choices made by the computer are not conducive to writing history, which as a field deals with gradations of reliability instead of an absolute truth.[1] It also

[1] Nick Tosh; John Tosh.

means that even as the LDA algorithm pulls apart texts and places an interpretative layer between the historian and the source, when it comes to making sense of the process, and interpreting the model *and* the text, we have to return *ad fontes* and look at the original source.[2] In presenting the way in which its topic modelling application works, this chapter generates new insights into the methodological considerations that are involved in using these tools as parts of historical research. Additionally, it will investigate if the design and production of a tool in this way is feasible for an individual researcher.

Topic Models: How Do They Work?

At the most basic level, a topic model is a way to sort text based on the words that occur within them. Imagine the desk of a scatter-brained professor, strewn with preparations for a conference paper they gave last summer, reports of department meetings, snippets of a book chapter they are working on, and pieces of a reader for the module they teach. A topic model performs the role of a secretary or research assistant tasked with systematically organising this chaotic jumble of papers. The academic has not explained to the research assistant how the documents should be grouped, but has given them a set number of unlabelled boxes into which all of the papers must be filed. The assistant would need to determine a strategy for grouping documents together based on their shared characteristics, and would then need to place each document in the correct box. This process would be shaped by the total number of boxes (or 'topics') available – a two-box sorting system (such as 'receipts' and 'research') would group together

[2] *Original Source* here means the version of the source that was fed into the LDA algorithm. Even if this is not its form prior to digitisation, it is the form it was in before the tool dissected it.

different documents than a five-box system ('conference', 'meetings', 'teaching', 'book', 'receipts'). A thousand-box system, on the other hand, might end up with every document in a unique box of its own. This, in simple terms, is the sorting process undertaken by a topic model — it takes a large set of unorganised documents, analyses their contents, and then groups each document into a predetermined number of topics (the boxes in our analogy) based on what it considers to be their shared linguistic properties.

While there are many variant implementations of topic modelling, they all base themselves on this same principle. To do this, topic models must make some assumptions. It assumes that each subject is discussed with its own language in its own text. It also assumes that the order of words in that text does not matter, but that occurring within the same text is enough to infer a relation between the words and the conceptual subject of the text. Each text, or document, is considered to be constructed out of a number of topics, as the author chose their words from the cluster(s) that relate to whatever it was they wished to discuss. These modelled topics are revealed by applying statistics to the text to infer the cluster from which the author chose their words.[3] Thus, an inferred or modelled topic represents a set of words which share a significant pattern of co-occurrence (two or more words that are used together in one text) or cross-occurrence (words that are not necessarily used together in one text, but which share a third word with which they are). For example, the concept 'Russia' represented by the words [Russian, Tsar,

[3] John W. Mohr and Petko Bogdanov, 'Introduction—Topic Models: What They Are and Why They Matter', *Poetics*, Topic Models and the Cultural Sciences, 41.6 (2013), 545–69 <https://doi.org/10.1016/j.poetic.2013.10.001>.

St. Petersburg] could be one topic, and the concept 'Navy' represented by the words [naval, fleet, battleship] another. The phrase "Russian battleship at St. Petersburg fleet review sinks" would therefore be classified as containing both of these topics; 29% of the phrase comes from topic one (Russian and St. Petersburg) and 29% comes from topic two (battleship and fleet). The other words in this sentence may belong to other conceptual topics with their own associated clusters of words.

It is important to note that at no point does the researcher interfere in the process of actually classifying the articles, or specify any a priori topics for generation. The classification relies on the comparison of word co-occurrence tables that the computer generates, which the researcher has no control over. Texts that get categorised into the same topic share some form of common meaning, as they share the same words. A single text may belong to more than one topic, if it shares commonality with more than one group of documents. This is where our analogy about the disorganised professor starts to become a bit stretched – the poor research assistant must now photocopy documents that could potentially be filed in multiple different boxes (should a hotel bill go in 'receipts' or 'conferences'?). Fortunately, for computers, this decision-making process is much less onerous.

We will continue to look in more detail at the topic modelling implementation in Latent Diriclet Allocation (LDA), which this thesis uses. LDA (and many other topic models) rely in large part on the representation of text as a bag-of-words. In this model, a text can be represented as a collection of word

frequencies, which are independent of each other. While this removes the positional and contextual information from the text, it means that it becomes possible to make Naïve (or Simple) Bayes assumptions about the text, and calculate the statistical likelihood that certain texts 'belong together' based on the frequency of certain words appearing. The Bag-of-Words model was developed in the 1960s as a way to facilitate the retrieval of documents, and was the foundation for some of the earlier spam filters in the 1990s.[4] It has been surpassed by more advanced classifiers, but it still enjoys popularity as it is fast and easy to implement.[5]

Latent Dirichlet Allocation functions by assuming that when the text was written, the author had a collections of bags of words, each of which they mentally correlated with a topic. It also assumes that the writer sought to produce sensible texts where words from the same topic-bag are used together more often than words from other topic-bags. We do not know the mental distribution of words over these bags, but we can use statistical analysis of the texts in the corpus to calculate the likelihoods for each word to belong to each topic. LDA does so by calculating the chance of one word belonging to a topic, then testing if this matches with any assignments it has made previously, averaging the distribution, and repeating this until all words are in the right topics. This above explanation is of course a simplification; the actual process is supported by a significant amount of

[4] David D. Lewis, 'Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval', in *Machine Learning: ECML-98*, ed. by Claire Nédellec and Céline Rouveirol, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer, 1998), pp. 4–15 (pp. 6–7) <https://doi.org/10.1007/BFb0026666>; Mehran Sahami and others, *A Bayesian Approach to Filtering Junk E-Mail*, AAAI Technical Report (Association for the Advancement of Artificial Intelligence, 1998), pp. 55–62 <https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-009.pdf>.

[5] Jason D. M. Rennie and others, 'Tackling the Poor Assumptions of Naive Bayes Text Classifiers', in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03 (Washington, DC, USA: AAAI Press, 2003), pp. 616–623.

mathematics. Readers who are not interested in this are free to move on to the next section.

Mathematics of LDA

LDA uses Bayesian statistics, which have their own underlying assumptions, key of which is that it allows the knowledge of the probability of an event occurring (such as a word appearing in a text) to be used to infer the probability of a related event occurring (the probability of the same word appearing in another text, for example). This assumption of interdependence can be problematic in other contexts, but when discussing language it is valid; no word is ever used in a vacuum, and texts are composed of interrelated words. The Bayesian foundations also allow these probabilities to be expressed as distributions, for example the often-used Gaussian curve (also known as Bell curve) or as any one of the other known distribution functions. Because we are dealing not with the probability of a single word, but of a range of multiple words with their own, but interconnected, probabilities, we can speak of a mixture distribution. Such a distribution essentially represents the weighted average of probability distributions for each of its constituents. A single text may thus be represented as a multivariate distribution of the probabilities that each of its individual words appears when picking a random word from it. However, these cannot be Gaussian distributions, as that is only defined for a univariate distribution.[6]

However, we are not dealing with single texts, but with a corpus of multiple texts, each represented by their own mixture distribution. Instead, each text

[6] Univariate distribution: probability distribution describing only a single variable.

represents a dimension in the probabilities space the text-mixtures occupy. Because of this multitude of dimensions involved, LDA has been designed to employ a Dirichlet distribution. This is a variant of the beta-distribution which allows for an unlimited number of variates and an unlimited number of dimensions. Unlike Gaussian distributions, which are defined by two terms (mean and deviation) Dirichlets are defined by a singular vector of positive, real numbers. For LDA, this is the vector derived from the observed word frequencies in the text.

But why undertake this statistical process? Again, the answer begins with Bayes: the goal of Bayesian statistics is to discover an unknown probability of a variable. But this variable itself can be unknown and unobservable. In such a case it is referred to as a latent variable. This, then is the goal of LDA: it uses a statistical analysis of the word Dirichlet distributions of the texts it is given to derive the unobservable, latent, topic distribution.

Drawing on de Finetti's proof that any collection of exchangeable random variables may be described as a distribution, Latent Dirichlet Allocation considers a document a mixture distribution of a known number of latent topics described by an unknown Dirichlet distribution.[7] Using Bayesian inference, it is possible to reconstruct the bags-of-words (topics) out of which the author drew for each component in the mixture distribution.[8] If α is the probability for a document to belong to a topic, and β is the probability distribution of each word to belong to

[7] Random Variables are mathematically defined as variables that have a probability (distribution).
[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3.Jan (2003), 993–1022 (pp. 994–98).

each topic, then the joint distribution θ (of which α and β are components) for *N* topics **z** and *N* words **w** can be found by solving:[9]

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(Z_n|\theta)p(W_n|Z_n, \beta)$$

The left side of this equation represents the complete topic model, the first probability factor on the right is the document assignment, which is itself composed of the product of the relative importance of each topic and all its constituent word-topic assignments. From this equation, we can derive the two steps that are required. The first is to determine the product of the word-topic probability; then this informs the probability of documents belonging to a topic. To solve this equation the algorithm goes through a series of steps:

- Initialisation, where the priors for the following steps are set. The probability for each word to belong to each topic is random, but around the range **w**/**z**. The likelihood of each document belonging to a topic is in the range 1/**z**. It is thus initially assumed that all topics are equally likely to occur and each word has an equal probability to belong to each topic.[10]

- For each word (W) in each document, it then calculates;

    1. How many words in this document are in each topic? (relates to β)

[9] Blei, Ng, and Jordan, 'Latent Dirichlet Allocation', p. 996 Eq. 2.
[10] A small amount of randomness is introduced to these initial probabilities to better reflect the realities of the assumed generative process of the documents.

2. How many documents in each topic Z have this word W in it? (relates to α)

3. What is the probability that W belongs to specific topic Z_n? (relates to θ)

4. It then updates the probability of an association between W and Z for every instance of W throughout the corpus.

- Repeat these steps until the Kullback-Leibner divergence in topic distribution $\theta = 0$, which means that the model has reached a 'steady-state' and further updates will not cause changes in the probabilities.[11]

Concretely, this means there are only two inputs that topic modelling programs need: a set *number* of topics assumed by the researcher (but not the *nature* of these topics), and lots of texts.

The typical output of the LDA algorithm will then consist of a list of tokens per topic, with a numerical value between 0 and 1. This value indicates the likelihood that this token appears in this topic. Even in perfectly seeded and homogenous corpora used for testing topic modelling tools this value rarely rises above 0.03. To give an idea of the minimal distances involved, in the LDA implementation of this topic, this association is calculated to nine decimal places accuracy. The main use of these association scores is that they can be used to judge how 'coherent' a topic is mathematically. However, this project will not be relying on these.

[11] Blei, Ng, and Jordan, 'Latent Dirichlet Allocation', p. 1005 Figure 6.

Crafting the Chisel: Topic Modelling

Despite the large variety of topic modelling programmes available, the most popular being the University of Massachusetts-Amherst's Machine Learning for Language Toolkit (MALLET), this project still produced its own topic modelling implementation, and argues that doing so is essential to using topic models as part of a historical methodology.[12] Some of the reasons to shun MALLET are purely practical; for example, it requires the different documents to be modelled to be in individual files containing their full text. While this is not a problem per se – the collated JSON format generated by the search queries could easily be automatically transformed into a collection of plaintext files – it would result in tens of thousands small files, which take longer to move to and from the external drive used to transport the data between the various locations than one big file. Additionally, the way that MALLET handles the importing of the text files means that it loads all of them in one go, and the amount of memory that is available for this is limited to 1 GB by default. This limits the size of the dataset we can use, and therefore constrains both the scope and nature of our research questions. While there are ways to increase the amount of memory that MALLET can use, these are roundabout

[12] McCallum, For examples of use see; Jobin Wilson, Santanu Chaudhury, and Brejesh Lall, 'Improving Collaborative Filtering Based Recommenders Using Topic Modelling', in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014, I, 340–46 <https://doi.org/10.1109/WI-IAT.2014.54>; Stefan Daume, Matthias Albert, and Klaus von Gadow, 'Assessing Citizen Science Opportunities in Forest Monitoring Using Probabilistic Topic Modelling', *Forest Ecosystems*, 1.1 (2014), 11 <https://doi.org/10.1186/s40663-014-0011-6>; Carina Jacobi, Wouter van Atteveldt, and Kasper Welbers, 'Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling', *Digital Journalism*, 4.1 (2016), 89–106 <https://doi.org/10.1080/21670811.2015.1093271>; Besnik Fetahu and others, 'A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles', in *The Semantic Web: Trends and Challenges*, ed. by Valentina Presutti and others, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2014), pp. 519–34 <https://doi.org/10.1007/978-3-319-07443-6_35>; Michael Röder and others, 'Detecting Similar Linked Datasets Using Topic Modelling', in *The Semantic Web. Latest Advances and New Domains*, ed. by Harald Sack and others, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2016), pp. 3–19 <https://doi.org/10.1007/978-3-319-34129-3_1>.

solutions, and do not address the fundamental issue, only treating the symptom. After a point, building one's own tool for the job-at-hand becomes easier than adapting someone else's tool to do that job.

However, there is a much more fundamental methodological reason for writing the tool. By not relying on the tools of others, but by building one themselves, the researcher needs to build a deep understanding of the processes and limitations that govern the application of the topic modelling program. Coding a tool requires the programmer to understand all the steps that that tool will take, know exactly what these steps do to the data, and have a reasonable understanding of why these steps work the way they do. By building the tool, the researcher becomes so familiar with it that they can also see where it might fail. In addition to this operational understanding, it allows the researcher to develop a tool that fits their material and their research question. Not all archives are created equally, and many have their own quirks, either created through decisions in the digitisation process or as an inheritance from their prior physical form; tools should take these into account. In a similar vein, research questions may necessitate different ways of analysing a tool's output. Being able to tailor the way the tool presents results to the research question that needs to be answered is a key advantage.

This thesis has made the case that crafting the topic modelling tool is advantageous, as it produces a tool that is tailored to the data in the archive and for the specific use case under investigation, which is methodologically and epistemologically transparent to the user and the reader. In light of this desire for transparency, the next section will discuss the development and operation of the topic modelling

tool this project assembled. It will specifically discuss the different parameters and settings that influence the modelling process and affect model quality. It will also comment on each step of the process as to whether it is feasible for an individual researcher to produce these kinds of tools.

Based on the search queries, discussed in the previous chapter, the server produced a series of json files, in which it collated the articles it had found. These files contained thousands of articles, often several hundred megabytes each, which were used to feed the topic modelling program. As one of the main objectives of this thesis is to explore the limits of what a single researcher can be expected to do themselves, and where they have to accept the use of pre-made tools or components for those tools; just as a carpenter doesn't make all their nails themselves when building a cabinet, relying on other people's work is inevitable when creating an analysis tool. This realisation strongly informed the choice for Python as the language in which to write the topic modelling tool, as it is a language that is common within the digital humanities and has many options to extend its operation using 'packages'. These packages are collections of community-produced, pre-defined methods and operations that can be included in a python build, which allows the coder to concern themselves with higher-level operations of the code instead of having to re-invent the wheel for a relatively common operation. This thesis relies on the Gensim topic modelling package, developed by Radim Řehůřek from 2008 onwards, which was the easiest to implement within

the skill threshold of a single researcher, and the computationally fastest implementation available.[13]

The topic modelling tool was written in python, and handles the translation from the corpus of documents represented in the json file to a topic model for analysis; its general workflow is represented by Figure 3.1. Initially, the sub-setted corpus of articles generated from the query is separated out into individual articles. Early on in the process, the corpus was batch-loaded: all the texts were loaded into memory at once before being processed. However, once the size of the corpus grew above 50,000 articles, issues began to occur with the amount of available RAM being insufficient to store all of them. The solution was to instead stream the corpus, which loads a single article, processes it into its less memory-intensive state, and then discards it.

[13] Radim Řehůřek and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta: ELRA, 2010), pp. 45–50.

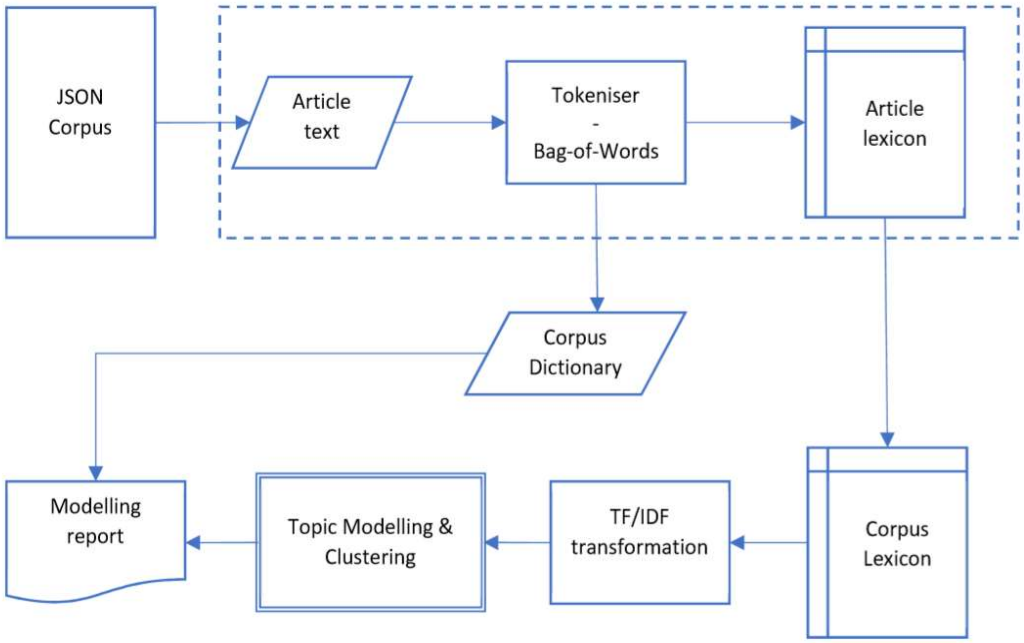


Figure 3.1 Schematic overview of the topic modelling process. The articles arrive as a single collated JSON file. Each article in this collection is tokenised and modelled as a Bag-of-Words, resulting in a lexicon of words and frequencies. Thus processed, the frequencies are modulated through a Term Frequency – Inverse Document Frequency step, before the LDA algorithm is ran on the resulting word-frequency matrix.

Data Pre-Processing

The articles are fed through a tokeniser, which separates words in a sentence and resolves hyphenation, and a translator that de-structures the text into a bag-of-words. The Bag-of-Words model significantly reduces the complexity of a text, as it retains only a Lexicon of the article; a list of the words in the text with its associated frequencies of occurrence.[14] This step also replaces the words in the text with numerical IDs to speed up processing. The word-ID relation is retained in the Gensim Dictionary and used at the end of the process to return models as words rather than the abstract IDs. There are a variety of tokenisers available for python, but this thesis used only two: a simple tokenise-on-space operation that was homebrew, and the Natural Language ToolKit package's advanced tokeniser and stemmer, which not only tokenises on the space, but also resolves hyphenation

[14] Lewis.

and deals with contractions ('you're' into the tokens 'you' and 'are') and verb forms ('does' and 'did' into 'doing'). In order to do this effectively, a spell corrector would have been needed to clean up the text before it was fed into the NLTK tokeniser, due to the large proportion of OCR errors in the newspaper transcriptions.[15] As will be discussed later, this was not used due to computing requirements and a simpler tokenise-on-space operation was used in nearly all cases.

The form of the Bag-of-Words model output is known as a lexicon: a table in which the frequency of occurrence of each word can be found. The tool generates these frequencies on an article level, while also replacing the word with a hexadecimal identifier, as these take less space to store than the word-strings. These article lexicons are then collated in a corpus lexicon, which records the term frequency per word per article. These raw frequency counts are then transformed using Term Frequency – Inverse Document Frequency (TF-IDF).[16] TF-IDF produces a relative appearance frequency by multiplying the term frequency, or how often a word appears out of all words in a document, with the number of documents in the corpus divided by the amount of documents that also contain

[15] Edward Loper and Steven Bird, 'NLTK: The Natural Language Toolkit', *ArXiv:Cs/0205028*, 2002 <http://arxiv.org/abs/cs/0205028> [accessed 3 March 2020]; Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (Sebastopol, CA: O'Reilly Media, Inc., 2009).

[16] Gerard Salton, 'Recent Trends in Automatic Information Retrieval', in *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '86 (Palazzo dei Congressi, Pisa, Italy: Association for Computing Machinery, 1986), pp. 1–10 <https://doi.org/10.1145/253168.253171>; Akiko Aizawa, 'An Information-Theoretic Perspective of Tf–Idf Measures', *Information Processing & Management*, 39.1 (2003), 45–65 <https://doi.org/10.1016/S0306-4573(02)00021-3>; Juan Enrique Ramos, 'Using TF-IDF to Determine Word Relevance in Document Queries', in *Proceedings of the First Instructional Conference on Machine Learning* (presented at the iCML-2003, Piscataway, NJ, 2003) <https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf>.

that word.[17] Once this step is completed, we run the LDA algorithm provided by the Gensim package on the resulting corpus.

Modelling Parameters

Several parameters influence the way the Gensim LDA implementation operates, which can be set during the initiation and modelling phases. This section intends to discuss these meta-parameters and the effect they have on the final model, as some of these variables can influence the final model greatly. Additionally, the way topic modelling tools work (through a tool built to control an underlying algorithm) means that they place two more layers between the reader and the text than there normally would be. This thesis feels it has to show transparency in this regard, as while the researcher knows exactly what all these buttons do, the reader does not.

It will, however, not discuss all the options available to the Gensim user: there are over twenty different variables to play with. However, some of these are only relevant for a distributed computing cluster, such as core allocation settings; or deal with technical behind-the-scenes aspects that have little to no impact on the model for the purposes of this thesis, such as the option to do calculations in 16, 32 of 64-bit integers, so these will not be discussed here. Only two of these variables were experimented with to any meaningful degree. The first was the number of topics (N_T), which shapes the entire topic modelling process. The

[17] Which can be written as: $tdf(t,d) = \frac{t_d}{L_d} * \frac{N}{df(t)}$, where we seek the frequency of term *t* in document *d* of length *L* in a corpus of size *N*.

others are the minimum required probability for text assignment, and dictionary lower- and upper bounds.

The choice of number of topics is crucial and has much more impact than any other variable. To return to our earlier analogy, the number of boxes into which our hypothetical research assistant organised the professor's disorganised documents. This is the key factor in shaping the results of a topic model, as it defines its granularity: a high number produces (theoretically) a large number of very specific topics, defined by the smallest variation in language; while a small number of topics produces more global collections that contain more documents. A too high number of topics can lead to overfitting, where only four or five texts are categorised in each topic – but they will be very clearly related. Too low a number will lead to the opposite, overstuffing, where a topic becomes so generic that it is impossible to classify. Striking the balance between these two extremes is dependent upon the intent of the researcher, but, crucially, also on the nature of the source material. Not all material is error-free enough to try and fit into a fine-meshed topic model. To return to our sculptor: attempting to carve Bernini's *David* out of pumice will just not work.

For digitised historical newspapers, there has been little structural enquiry into the ranges of N_T that produce useful results for this kind of historical study. Thus, this thesis set out to experimentally verify the upper and lower limits for N_T by generating and analysing a series of topic models until it establishes which produce meaningful topics. In other words, should we group articles into a relatively small number of topics, a large number of topics, or something in

between? And what effect might this decision have on the questions we ask of our dataset? For the initial topic models the N_T -values used were 10, 100, and 150, in keeping with the literature available for topic modelling modern corpora of equivalent size, with $N_T = 10$ intended to establish a bottom line.[18]

Evaluation of topic model topics is a persistent problem that has not been solved yet, but in order to verify the success of the various numbers of topics, we need a judgement on 'usefulness'.[19] While there are mathematical measurements like Perplexity to determine how well a topic model preforms, these do not correspond to how coherent a human finds a topic.[20] This thesis tried two ways to determine the usefulness of a topic model: one experimental and one more humanistic. For the first, this project followed the work of Chang et al, who developed topic intrusion tests.[21] At random, a pair of documents would be added to a topic in a topic model with $N_T = 20$ around the keyword 'India'; if the human user could spot the intrusion, the topic was coherent. However, this approach did not produce satisfactory results. As the coherence of the topics was generally poor to begin with, intrusion tests only exacerbated the issue. In one notable instance,

[18] Blei, Ng, and Jordan, 'Latent Dirichlet Allocation', p. 1010; Chang and others, p. 293.
[19] Leon van Wissen, 'Topic Modelling "De Gids": An Explorative Study into the Use of Topic Modelling on a Cultural Periodical' (unpublished Research Master Thesis, Vrije Universiteit, 2019), pp. 66–67 <https://www.leonvanwissen.nl/publication/vanwissen-2019-degids/vanwissen-2019-degids.pdf>.
[20] A. De Waal and E. Barnard, 'Evaluating Topic Models with Stability', in *Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa* (presented at the PRASA 2008, Cape Town, South Africa, 2008), pp. 79–84; Chang and others; David Mimno and others, 'Optimizing Semantic Coherence in Topic Models', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11 (Edinburgh, United Kingdom: Association for Computational Linguistics, 2011), pp. 262–272; Arthur Asuncion and others, 'On Smoothing and Inference for Topic Models', *ArXiv:1205.2662 [Cs, Stat]*, 2012 <http://arxiv.org/abs/1205.2662> [accessed 3 March 2020]; Keith Stevens and others, 'Exploring Topic Coherence over Many Models and Many Topics', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12 (Jeju Island, Korea: Association for Computational Linguistics, 2012), pp. 952–961.
[21] Chang and others, p. 291.

the pair of intruding documents were more coherent than the remainder of the topic.

Thus, the thesis turned to its humanistic roots and chose to rely on a close reading of the ten best-fitting texts for each topic. If it could determine a coherent narrative between the texts, no matter how broad or specific, the topic was judged to have 'meaning'. The advantage of this approach is that it closely mirrors the way a humanities researcher would work without digital tools. One way in which a historian may attempt to make sense of an archival collection is by constructing an (internal) narrative around the sources, to find some form of coherence.[22] This way of evaluating the validity of the topics proved to be much more valuable than binary intrusion testing, as it did not discard completely topics that were only partially coherent. These could still be read, understood and, with appropriate epistemological scepticism and caution, enhance our understanding of the corpus. They also further understanding of topic models themselves, and this thesis can report that not only do words that share semantic meaning form topics together, so do certain kinds of transcription errors. Form this, we deduced that certain kinds of transcription errors are more likely in certain kinds of content. For example, railway timetables are generally badly transcribed, but still form a coherent topic together because the transcription fails in identical ways.

Having produced the three models with varying numbers of topics, it was found that the 100- and 150-topic models were meaningless. For the 150-topic

[22] W. H. Dray, 'On the Nature and Role of Narrative in Historiography', *History and Theory*, 10.2 (1971), 153–71 <https://doi.org/10.2307/2504290>; Allan Megill, 'Recounting the Past: "Description," Explanation, and Narrative in Historiography', *The American Historical Review*, 94.3 (1989), 627–53 <https://doi.org/10.1086/ahr/94.3.627>; Breisach, pp. 380–82.

model, only 23 topics (15%) were coherent and could be identified; 82 of the topics in this model shared indicative words. On the 100-topic model, this rose to 33 topics, while 59 shared one or more indicative terms. Many topics were differentiated by variants of OCR errors, which were easily recognisable, as they contained only a dozen or so articles, and often had indicative tokens in common with all other topics. The main indicator however was the lack of coherence in indicative tokens. For indicative tokens, coherence was considered not just between the words, but also with the texts in the topic; a topic where the indicative words form a narrative of fishing and seamanship while the texts form a coherent collection on cattle farming in Cornwall is not a coherent and sane topic. Therefore, the number of topics was reduced drastically, stepping down to 50, 30 and finally 20 topics, which produced much more sane models. The final choice for 20-topic models was due to their good balance between creating sufficient distinctions in the corpus and production of coherent topics. On average, in the 20-topic model 16.1 topics (80.8%) were coherent. A comparison of a topic on these models is offered in Figure 3. for the 150-topic model and Figure 3. for the 20-topic model.

Topic 8/150 – 21 articles
0.004*"chatham." + 0.004*"bill." + 0.003*"speech" + 0.003*"scene" + 0.003*"threatened" + 0.002*"burial" + 0.002*"legislative" + 0.002*"bill" + 0.002*"prices" + 0.002*"last." + 0.002*"bengal," + 0.002*"assembly"
0) FOREIGN. FRANCE. PARIS, MONDAY
1) LATEST INTELLIGENCE. FRIDAY EVE.ING I. THE QUEEN'S HEALT
2) ARRIVAL OF A NEPAULESE EMBASSY AT SOUTHAMPTON
3) MISCELLANEOUS NEWS
4) THE WEAVERS' STRUGGLE
5) STATIONS OF T1HE BRITISH ARMY
6) EXTAORDINARY SCENE IN ST, MICHAEL'S CHURCH, LIVERPOOL
7) MARKETS. LIVERPOOL PROVISION MARKET
8) MARKETS. MANCHESTER HAY Axo STRAW MARKET
9) LECTURES TO TlE _WORKING C'LIMSSE

Figure 3.2 Typical topic (Indicative tokens and article titles) in a 150-topic model. Note that 'bill' and 'bill.' appeared in 82 of these topics as indicative token. This topic has no clearly definable meaning.

Topic 15/20 – 4487 articles
0.012*"do." + 0.011*"ditto" + 0.011*"wheat" + 0.008*"barley" + 0.008*"market" + 0.008*"ditto," + 0.007*"qr." + 0.007*"oats" + 0.007*"qrs." + 0.007*"supply" + 0.006*"demand" + 0.006*"white"
CORNWALL MARKETS
Idem. (3 times)
LONDON AND COUNTRY MARKETS
LONDON MARKETS
Idem.
WAKEFIELD CORN-EXCHANGE
Idem. (2 times)

Figure 3.3 Typical topic (Indicative tokens and article titles) of a 20-topic model. Note the clear cohesion between the articles and indicative tokens. This topic clearly relates to the grain markets.

Figure 3.2 shows an incoherent topic model. It is small, which is not a classifier for incoherence, but it is a warning sign. A topic of only 21 articles should be extremely clear – after all, out of thousands of articles *only* these 21 were considered to share something by the algorithm. While some of the texts share features, such as a place (6 and 7, Liverpool), content (7 and 8, Markets), or a theme (4 and 9, the working class), there is no shared common ground between all of them. The same can be seen in the indicative words: some can be matched to an individual text, such as text 2 with 'bengal', there is no overarching theme or connection. The topic in figure 3.3 from a 20-topic model on the other hand shows

a high degree of coherence. It contains several articles from the same weekly column, and there is a clear connection between the texts and the indicative terms.

Yet this was not always the case; for some keywords and time periods, the 20-topic model did not perform as well as 50- or 100 topic models. However, in order to be able to make a meaningful comparison between models, the 20-topics were retained, even if they had higher rates to incoherent topics. This points to the conclusion this thesis presents on number of topic section when modelling historical corpora: the researcher needs to determine the optimal number of topics themselves as fits their data and research question, rather than relying on the recommendations of other researchers.

One other value this thesis experimented with was the minimum required probabilities for text assignment. The first, minimum text-topic association probability (φ_{min}), governs the threshold at which a text is assigned to a topic.[23] If this is set to zero, a text will be assigned to any topic that contains a word that is also within it; if a text has one mention of India amongst thousands of words, it will be considered part of the topic that is assigned the word 'India'. Conversely, if its set to one, only a text that is entirely composed of 'India'-topic words will be assigned to that topic. By increasing this value from the default (0.01), varying between 0.05 and 0.1, it was found that topic cohesion improved slightly. This improvement was hardly noticeable on the 20-topic models that were already coherent before the change, but on those selected keywords and time periods

[23] Pronounced Phi-min.

mentioned above, where higher topic numbers would have been better, increasing φ_{min} offset some of the incoherence.

Finally, the modelling experimented with different lower and upper bounds for the dictionary. These values change how the Bag-of-Words is generated from the original texts, which influences the resulting topic model. In the Bag-of-Words step, at the beginning of the process, the texts are transformed into frequency counts, and words are discarded based on their frequency of appearance and these parameters. The lower bound is the minimum number of times a word must appear to be retained. This is designed to filter out extremely rare words, for which an accurate topic assignment cannot be inferred, because it only appears in one or two documents. In the case of this thesis this was raised from the default of 5 to 2, as a way to retain meaning that would otherwise be filtered out due to OCR errors. The upper filter bound filters out words that are so common in the corpus, that they can equally be assigned to every topic, but which are not included explicitly in the list of stop words. It rejects words that appear in more than a set percentage of the corpus; by default, this value is 50%, but in response to OCR errors, and the desire to remove as little potentially indicative words as possible, this was lowered to 33%. These settings produced somewhat more coherent topics on the error-riddled text in the archive.

These findings on the significance of these two sets of variables highlight two important points. First, the importance of the researcher to be able to fine-tune the topic modelling to suit the data it operates on, which is only possible by building the tools used oneself. Second, that topic models are very sensitive to

minor changes in the modelling parameters, and should thus be used with caution by historians, as just a small change in an otherwise hidden variable can cause changes in the quality of the resulting model. Finally, the choice of the number of topics, as the most important factor determining the model's form, should always be tailored to the data and research question.

Model Viewing

Having established the parameters that influence the modelling, we can now turn our eyes to the results of the modelling cycle, and the way these results are analysed. This section will introduce the ways in which the output of the topic model is transformed from its default form of a list of terms to a functional output that can be used for historical research. The viewing and analysis aspects in this discussion are interrelated, as this project developed its own way to view the topic that was heavily informed by my own research practices. Thus, issues in the analysis, for example due to missing information about an aspect of the topic model resulted in these features being added to the viewer.

The default form of a topic model in Gensim is illustrated in Figure 3.4 (top). In no particular order, it produces a list of the topic IDs and the ten most indicative words for that topic. This is by no means ideal for understanding a topic, as we have no idea which texts informed the creation of the topic, and we do not know how large the topic is compared to the corpus. This is not a novel problem, and a variety of researchers have developed topic visualisation tools and techniques. Van Wissen, for example, uses word squares sized to the importance

of each word.[24] Others have used bar graphs, cluster visualisations, and all manner of other graphical aspects to improve understanding.[25] Yet none of these were still available or compatible with my system. One tool that was seriously considered was DFR-browser developed by Andrew Goldstone for JSTOR's *Data for Research* Metadata repository, however, this tool was written in an incompatible language (Java), and would only work out of the box with MALLET.[26] Again, as with the choice not to adopt MALLET, the potential time investment to alter such an existing package to work with what this tool does, and then alter it to do what fits my research practices would be prohibitive. The more practical pyLDAvis package was rejected, as while it would be easy to implement, it only shows the features of the topic model, and does not allow for the original text to reappear.[27] Instead, I developed my own way to view a topic model, seen in Figure 3.4 (bottom).

The first goal of the topic discovery process is to determine what a topic means. This assumes that a topic is coherent, and that its meaning can be made sense of by a human reader. There are two possible avenues to discovering this meaning: through the list of indicative words, or through the texts belonging to the

---

[24] van Wissen, p. 78.

[25] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda, 'Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents', in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08 (Las Vegas, Nevada, USA: Association for Computing Machinery, 2008), pp. 363–371 <https://doi.org/10.1145/1401890.1401937>; Allison June-Barlow Chaney and David M. Blei, 'Visualizing Topic Models', in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012 <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4645> [accessed 4 March 2020]; Jason Chuang, Christopher D. Manning, and Jeffrey Heer, 'Termite: Visualization Techniques for Assessing Textual Topic Models', in *Proceedings of the International Working Conference on Advanced Visual Interfaces* (ACM, 2012), pp. 74–77; Jaimie Murdock and Colin Allen, 'Visualization Techniques for Topic Model Checking', in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015 <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10007> [accessed 4 March 2020].

[26] Andrew Goldstone, *Dfr-Browser*, version 0.8.1, 2019 <https://github.com/agoldst/dfr-browser> [accessed 4 March 2020].

[27] Ben Mabey, *PyLDAvis*, version 2.12, 2020 <https://github.com/bmabey/pyLDAvis> [accessed 4 March 2020].

topic. Both of these have their downsides: given the scale at which this tool operates, there are inevitably too many texts within a topic to read, while the nuance of a topic can be difficult to discover through a list of words alone. Thus, this thesis developed the concept of a 'topic window'. From the texts associated with the topic, the twenty texts that are most representative of the topic based on their text-topic association scores are selected and presented to the researcher in full. These offer a 'window' through which we can look at what is inside the topic, in the assumption that the general connection between these texts is representative for the relationships between all texts in this topic.

The second question when understanding a topic is, "how important is this topic to the corpus?" When answering a historical research question, the quantity of evidence is important, as a handful of sources supporting a thesis will make it difficult to make a claim to ubiquity. Thus, with each topic, the viewer shows the percentage of the corpus that this topic contributes to. It should be noted that the percentage calculated is an approximation, arrived at by taking the sum of the article-topic match percentage. This allows for an article that has a 50% match with two topics to contribute to both those topic's sizes in the calculation over the corpus. There are however two important caveats to keep in mind when reviewing the percentage-of-corpus statistic. First, a topic which has many articles of low certainty of matching will have a lower indicated relative size than a topic consisting of a few articles with very sure matches. Second, it means that the percentages of relative size of the topic in the corpus do not add up to 100%. While this is

counterintuitive at first, this is the recommended way for calculating topic size relative to the corpus in Gensim.[28]

Model Analysis

Having developed a way to view the output of the topic model, we may now start the analysis and annotation of the topic models. The goal of the analysis is to ascribe meaning to the topics in such a way that they can be compared between models, to show change over time or difference between two sets of keywords, which necessitated certain decisions in the accuracy of topic labelling. This section will describe the way in which the analysis of the topic models takes place, and discuss the implications of decisions taken during analysis, such as the definition of categories for comparing topic models. It will also discuss the way topics that could not be identified or which were incoherent were handled.

Before the models can be analysed, the method by which they are interpreted and given meaning needs to be developed. As the goal of this project is to develop a way to use topic models historically, it needs to model its analysis on historiographical praxis. My belief was that this goal would best be served by retaining as much of the historian's craft in the process as possible, and treating a topic in a generated model as a source cum archival collection. This is another key reason for not adopting an existing topic model viewer, as discussed above: none of the available options linked back to the 'original' article text.

[28] Radim Rehurek, 'LDA Corpus Topic Composition - Google Groups', *Google Groups*, 2014 <https://groups.google.com/forum/#!searchin/gensim/topic$20percentage$20corpus%7Csort:date/gensim/3cmG23E4Wl4/1zhxiT1d8EIJ> [accessed 2 July 2017].

```
Dictionary(38332 unique tokens: ['gentleman.', 'prnssians', 'herald,', 'hazards', 'roumnania']...)
(0, '0.135*"WAS" + 0.126*"IT" + 0.125*"IS" + 0.121*"AS" + 0.118*"NOT" + 0.116*"WOULD" + 0.112*"WHICH" + 0.112*"T" + 0.110*"THAT" + 0.109*"ARE"')
(1, '0.326*"SERVIAN" + 0.267*"TURKISH" + 0.174*"JULY" + 0.167*"TURKS" + -0.158*"LORD" + 0.137*"PRINCE" + -0.124*"AMEER" + 0.116*"SERVIANS" + -0.114*"WOULD" + -0.109*"POLICY"')
(2, '0.390*"T" + 0.292*"\\\'" + 0.287*"E" + 0.228*"R" + 0.212*"C" + 0.204*"*" + 0.189*"S" + 0.176*"N" + 0.172*"1" + 0.167*"D"')
(3, '-0.460*"FRENCH" + 0.278*"SERVIAN" + -0.214*"PRUSSIAN" + 0.213*"RUSSIAN" + 0.210*"TURKISH" + -0.189*"FRANCE" + -0.186*"GERMAN" + -0.132*"PRUSSIANS" + 0.129*"RUSSIA" + -0.120*"PARIS"')
(4, '-0.457*"GREEK" + 0.449*"SERVIAN" + -0.185*"RUSSIAN" + -0.163*"APRIL" + 0.150*"SERVIANS" + -0.130*"CONGRESS" + 0.120*"TURKS" + -0.105*"PORTE" + -0.101*"TREATY" + -0.097*"NEW"')
(5, '-0.351*"GREEK" + 0.290*"AMEER" + 0.219*"BRITISH" + 0.188*"AFGHAN" + 0.150*"ALI" + 0.147*"J" + 0.139*"CABUL" + -0.131*"SERVIAN" + 0.127*"—" + 0.126*"RUSSIAN"')
(6, '0.383*"J" + 0.351*"\\\'" + 0.236*"*" + 0.221*"—" + 0.203*"■" + 0.203*"•" + 0.201*"SERVIAN" + -0.198*"E" + -0.183*"T" + 0.176*";"')
(7, '0.407*"RUSSIAN" + -0.256*"SERVIAN" + -0.250*"GREEK" + -0.187*"LORD" + 0.180*"RUSSIA" + 0.169*"RUSSIANS" + -0.154*"WAS" + -0.132*"JULY" + -0.127*"SIR" + -0.126*"HOUSE"')
(8, '-0.679*"GREEK" + 0.142*"APRIL" + -0.120*"AMEER" + -0.101*"BRITISH" + -0.100*"FRENCH" + 0.100*"LORD" + -0.098*"AFGHAN" + -0.098*"ALI" + 0.097*"RUSSIA" + -0.093*"TROOPS"')
(9, '0.323*"FRENCH" + 0.254*"AMEER" + 0.210*"SERVIAN" + 0.190*"JULY" + -0.161*"APRIL" + 0.159*"PRUSSIAN" + 0.154*"BRITISH" + 0.143*"FRANCE" + 0.141*"GERMAN" + -0.127*"CARLIST"')
(10, '0.257*"AUSTRIAN" + -0.230*"APRIL" + 0.211*"SERVIAN" + -0.210*"RUSSIAN" + -0.197*"JULY" + -0.173*"LORD" + 0.169*"CARLIST" + -0.166*"RUSSIANS" + 0.150*"IS" + -0.131*"WAS"')
(11, '-0.493*"APRIL" + -0.407*"CARLIST" + -0.274*"CARLISTS" + 0.273*"AUSTRIAN" + -0.192*"DON" + -0.150*"SERVIAN" + -0.137*"SPANISH" + 0.128*"AUGUST" + -0.123*"MADRID," + -0.112*"CARLOS"')
(12, '0.542*"JULY" + -0.353*"SERVIAN" + -0.317*"APRIL" + 0.251*"AUSTRIAN" + 0.244*"CARLIST" + 0.152*"CARLISTS" + 0.109*"DON" + -0.102*"—" + 0.102*"AUGUST" + -0.101*"STEAMER"')
(13, '0.496*"JULY" + -0.416*"AUGUST" + -0.194*"AUSTRIAN" + 0.152*"—" + -0.139*"INSURGENTS" + 0.130*"CONGRESS" + -0.113*"RUSSIAN" + 0.110*"SHALL" + 0.108*"NEW" + 0.102*"STEAMER"')
(14, '0.529*"APRIL" + -0.382*"RUSSIAN" + 0.339*"AUSTRIAN" + -0.155*"SERVIAN" + 0.123*"PORTE" + 0.118*"INSURGENTS" + -0.113*"CARLIST" + -0.108*"RUSSIANS" + 0.104*"—" + -0.101*"CHINESE"')
(15, '0.461*"AUGUST" + 0.327*"JULY" + -0.154*"CARLIST" + 0.141*";" + 0.131*"(FROM" + -0.129*"CONGRESS" + 0.127*"TELEGRAPH.)" + 0.125*"PER" + 0.124*"SUBMARINE" + 0.115*"APRIL"')
(16, '0.488*"AUGUST" + -0.463*"AUSTRIAN" + 0.185*"SHALL" + 0.180*"CONGRESS" + -0.144*"APRIL" + 0.140*"PORTE" + -0.140*"JULY" + 0.137*"TREATY" + -0.125*"RUSSIAN" + -0.106*"GERMAN"')
(17, '-0.373*"AUGUST" + 0.218*"OCT." + 0.212*"(FROM" + 0.201*"TELEGRAPH.)" + 0.194*"SUBMARINE" + -0.184*"AUSTRIAN" + -0.176*"APRIL" + 0.175*"(BY" + -0.173*"RUSSIAN" + -0.168*"CAPE"')
(18, '-0.341*"J" + 0.322*"FENIANS" + 0.321*"FENIAN" + 0.281*"*" + 0.202*"UNITED" + -0.167*"|" + 0.158*"CANADIAN" + 0.152*"\\\'" + 0.151*"STATES" + -0.142*"CAPE"')
(19, '0.405*"—" + -0.319*"*" + -0.256*"\\\'" + 0.252*"J" + -0.222*"CAPE" + 0.204*"AUSTRIAN" + 0.132*"|" + -0.132*"(FROM" + 0.130*"YOU" + 0.122*"T"')
```

```
360 ==========================================================================================================================================
361 topic no: 10     perc of corpus: 0.003802
362 Topic Stats:
363 0.180*"lung" + 0.013*"johnny" + 0.012*"1894." + 0.010*"beware" + 0.008*"team" + 0.008*"promotes" + 0.007*"sat.," + 0.006*"rheumatism." + 0.005*"bristol" + 0.005*"parcel" + 0.004*"match" + 0.004*"bales."
364 ------------------------------------------------------------------------------------------------------------------------------------------
365 # 0 top article ID: HUCE-1887-11-19-0002-003     top article score: 0.913636
366 IK
367 CRICKET AUSTRALIA. Sydney. Saturday.—A match between Shaw, Shews- burr and LUly white's team end a Sydney eleven resulted in-day In tho defeat of the Englishmen by 10 wlckeia. An Australian cricket team will visit England next
    summer. M___BoDB_n*, Tuesday.—The match between Mr Vm-ion's Eleven and Twenty*two of Oaatlemalne ended in _^dnw In favour of the Englishmen. MatBOOBJ-**, Thuraday.— The match between Vernon's Eleven aad Eighteen of Sandhurst
    terminated in a dnw.
368 ------------------------------------------------------------------------------------------------------------------------------------------
369 # 1 top article ID: HUCE-1887-12-05-0004-026     top article score: 0.894444
370 CRICKET IN AUSTRALIA.
371 Bbisbakb, Saturday.— Tho Sbatr. Shrewsbury. __,• Llllywbite team commenced a match yesterday ag-inn Eleven of Brisbane. In tbe first Innings the Eoejii.. men made IJ3 and the home team 93. In the s-c iz i innings the Eagllahmsn
    scores 152 for £.ye wickets. Stdmet, Sunday.— The match batween Veraeai Eleven and Eighteen of Paramatta b<.s ended in a etas
372 ------------------------------------------------------------------------------------------------------------------------------------------
373 # 2 top article ID: HUCE-1882-02-18-0007-043     top article score: 0.894444
374 CRICKET IN AUSTRALIA
375 The great matoh ef the English tour ln Autralla was played at Melbourne on Deoember 31st, January 2nd, 3rd, 4th, and Sth, between Shaw. Team and the com- bined Eleven for the Australian Colonies. Spofforth and Allan did not
    play, but the Australian Eleven waa much stronger than that whioh played Eogland at tho Oval in 1880. The Englishmen scored 294 and 308. and Au.tralla 320 and 127 for three wickets ; the game being then left drawn, after nearly
    four days play. T. Horan scored the highest for Australia, and Selby for Eogland.
376 ------------------------------------------------------------------------------------------------------------------------------------------
377 # 3 top article ID: HUCE-1882-12-06-0003-013     top article score: 0.864286
378 THE ENGLISH CRICKETERS IN AUSTRALIA.
379 At Sydney on Monday, a match was concluded, after three day*,' play, between tbe Hon. Ivoßligh's team and an eleven of New South Wales. Delightful weather prevailed throughout the contest, whioh was witnessed by many thousands
    of spectators. The English team were iv grand forru.-and made 461 in their first innings, which included a faultlessly played 144 by C. H. F. Leslie The Australians scored 152 in their first innings and 165 iv their necond, s >
    tnat the visitors wen the m&'.ch by an innings and 144 ran.
380 ------------------------------------------------------------------------------------------------------------------------------------------
381 # 4 top article ID: HUCE-1879-03-01-0003-021     top article score: 0.864286
382 THE ENGLISH CRICKETERS IN AUSTRALIA.
383 The Sportsman of Wednesday morning contains the following telegram, dated Melbourne, February 25, 5*38 p.m., from one of toe English Eleven : — " This afternoon we brought to a conclusion our eighth match. Our oppo- nents on
    this occasion were an eleven of Victoria. We were the first to bat, and the result of our Innings was the large total of 325. To this the Victorians replied with 261, so that we were 64 runs to the good when we went In a second
    time. In our next c? say, however, we were far less fortunate, falling to reach our previous score by 154 runs, our total being 171 runs. The Colo- nials, batting in very good form, were quite equal to the task of not only
    pulling off their arrears, but of winning, as they put together the requisite of 236 runs with the loss of eight wickets, so that we were defeated, after an exciting game, by only two wickets.
384 ------------------------------------------------------------------------------------------------------------------------------------------
```

Figure 3.4 Example of default Gensim output to terminal (top) and self-developed output using Notepad++ (below). The former only provides the words and their relative weights from which a meaning has to be constructed. The latter offers additional information, such as the relative size of the topics, the original texts, and the degree by which those texts associate with the topic.

But this connection is crucial and the cornerstone of the use of topic models when writing history. The indicative terms of a topic signpost discovery, but the actual nuance can only be gotten from the text itself by a close reading. A 'window' into the topic therefore has to show the source's text.

Through close reading of the window of topic-matching articles, the historian has to ascribe a 'meaning' to the topic. To return once again to our prior analogy, the professor now returns to their office and finds that their paperwork has been dutifully organised into boxes – but the boxes do not have any labels. Just as the professor now needs to identify what each box contains, a historian analysing a topic model must determine what the texts grouped into a particular topic have in common. We call this the 'lowest common textual denominator', as the challenge is to be as specific as possible when undertaking this process. This always takes place in dialogue with the other topics in the model: if two topics appear similar at first glance, the interpreter has to look closer and investigate if it is possible to work out what caused them to appear different to the algorithm and what the semantic meaning of this difference is. These commonalities can be very diverse and nebulous at times. In some cases, elements of the layout of an article, such as the large amount of semicolons in lists, causes the topic to stand out. In other cases, the communal factor is the language; Welsh newspapers cluster together into their own topic, which sometimes splits into a 'general news' and 'adverts, trade and tables'. These meanings are then recorded first in the text file of the topic model's output, and later in an excel worksheet.

For the purposes of this thesis, an attempt was made at retaining a somewhat standardised classification system of the topics identified. This evolved out of the observation that sometimes, topics were so specific to a keyword, time period or newspaper that comparison with other topic models was difficult based on the most-specific topic labels. Therefore, seven categories were used to generalise the topic labels. These were *economy*, for topics touching on the financial market and national wealth; *trade*, for topics relating to goods and produce prices and local economic life; *politics*, which includes all news from Parliament and government, both foreign and domestic; crime, which contains the police court topics; *welsh news*, which collects all topics generated because of the inclusion of welsh papers in the model; *adverts*; and *personal news*, which forms an eclectic category of birth, death, and marriage notices, honours bestowed by the queen and published in the London gazette, and court and society gossip. An eighth category, *miscellaneous* was used to categorise all topics that did not fit any of the above.

These are extremely broad categories, as when comparing topic models, the category labels can only be as good as the least specific topic. While this ability to compare topic categories between different models is a major advantage, it also means that while the topic models presented here are very good at providing a global overview of trends in the press, they lack the ability to reliably do so in a nuanced way. This is due to the models only being able to define topics or categories that are significantly different from each other in the kinds of language that they use. The formation of these categories is down to them having a jargon or specialised vocabulary that is only present in these articles. For example, market bulletins are the only place where 'ditto' consistently appears, and 'bill' and

'member' are largely unique to parliamentary reports. Yet while the various noble ranks are indicators for court and society news, they are not unique enough and appear in such a variety of documents that we cannot form a 'courtly' category, but instead have to generalise that as part of 'personal and society news'. This means using topic modelling on datasets of this size and diversity can only make global comparisons between different sub-corpora, which limits it to research questions about the contexts in which certain terms appeared.

The assumption that all topics are coherent *and* meaningful does not hold true for every topic in every topic model. Sometimes, even using twenty topics, some overfitting still occurred. These topics were identifiable by representing very small slices of the corpus (less than 1%), and have one or two texts with a strong association to the topic (greater than 95%), while the remainder of the texts are only lightly associated. This thesis also identified one particular way in which the model may produce an incoherent model that is still meaningful. This is what it terms a 'bit topic', which represents a more nebulous concept, for example a style of writing, that only really shows in conjunction with other topics. These topics are often recognisable by not only making up a small percentage of the corpus, but also by the documents within them having low scores within that topic. For example, when applying these topics to articles mentioning America, one of the topics produced dealt with a variety of disasters and atrocities, all written in very evocative and emotional prose, such as 'THE MASSACRE OF MISSIONARIES', 'DEATH IN A SNOWDRIFT' and 'THE PERILOUS STATE OF THE

ATLANTIC'.[29] Topics such as these, as well as any overfit topics, were given their own category: unclassified, and were excluded from further analysis for the purposes of this particular project. This was done with a realisation that while they are valuable for research questions looking at emotional language in the English press, they are not useful for research questions within the theoretical framework of this project. This ties in with the point this thesis makes about the adaptation of the tool to the research question.

After the labelling of the topics has taken place and has been recorded in an excel worksheet, statistical and graphical analysis takes place. This depends greatly on the research question that is to be answered, but as a minimum I would advise generating a stacked bar graph of categories to compare the composition of the generated model against a 'baseline' topic model. This format is easily extended to temporal topic models. At this point, it also becomes clear how much of the corpus has been classified, and how much is unaccounted for in the unclassified category. While each researcher should set their own standards for what level of unclassified material they are willing to accept, this project used 65% as its standard. That is, at least 65% of the corpus had to be successfully classified by the model in order for it to be considered sound.

In conclusion, the analysis process is patterned on the process of historical discovery, and consists of a close reading and annotation phase, and a statistical phase. In the process of the first, the topics in the generated topic model are

[29] 'Death in a Snowdrift', *Pall Mall Gazette* (London, 18 September 1890), p. 6; 'The Perilous State of the Atlantic', *Pall Mall Gazette* (London, 21 September 1894), p. 8; 'The Massacre of Missionaries', *Pall Mall Gazette* (London, 21 September 1894), p. 6.

labelled with a number of categories, which cover the entire research project. These categories can only be as specific as the least detailed topic, and are thus strongly informed by the quality and size of the dataset. The statistical analysis is highly dependent on the research question, and offers a quantitative comparison between models. My recommendation is to use these topic models in an exploratory context for historical research. This means that the goal of the user is not to generate the most mathematically perfect clustering, but rather to generate one that is easiest to interpret. The end result of the modelling phase should not be an all-explaining model, but rather a collection of documents, organised by the topic model, that fuels further questions on behalf of the researcher.

This means that a model that contains topics that cannot be classified, or topics that are flat-out erroneous (either through overfitting or overstuffing) need not be discarded in its entirety. As historians, we have to treat a topic model like we would any other piece of historical evidence, and critique it to at least the same standards. Even if its composition falls below the threshold of certainty that means it cannot be used as the foundation for a claim to truth per se, it can, and should, inform further research questions, and serve as contextual evidence. Topic models can gesture towards an interesting research question, or even towards a solution to a research question, but it is incapable of generating an objective truth on demand which stands above scrutiny.

Evaluation

While overall the workflow described above worked well and produced useful results, the experience of modelling the texts was filled with trial and error, and

there were several paths taken that turned out to be dead ends and wastes of time. Thus, having shown the process of generating and analysing the topic models, I will now discuss my experiences with the topic modelling tool, the way its limitations shaped the research project, and make recommendations for future improvements. This evaluation will cover three key areas: the influence of data quality on model quality; the Analysis Window, for which several improvements are proposed; and the computational performance of the topic modelling tool, which has implications for the opportunities for a single researcher to use these tools, and the way they can function in a collaborative environment with computer scientists.

Data Quality

The first of these deals with the limitations of the topic model, from both a technical and an epistemological point of view. As is the case with any form of data science or statistical analysis, the results can only be as good as the quality of the data that is supplied. This is known as the GIGO-principle: Garbage In, Garbage Out. In the case of topic models, this is in full force when it comes to the quality of the OCR-transcribed texts used to build the models, which were often poor and filled with errors. This meant that the models essentially had to categorise and classify two different languages within the same document: regular English or Welsh, and English or Welsh as misspelled by the OCR transcription. This runs counter to one of the base assumptions of topic models: that each document is a self-contained capsule of language, which uses that language to discuss a singular issue, and thus the topic models generated don't reach the levels of accuracy that research in computer science or computational linguistics would suggest the

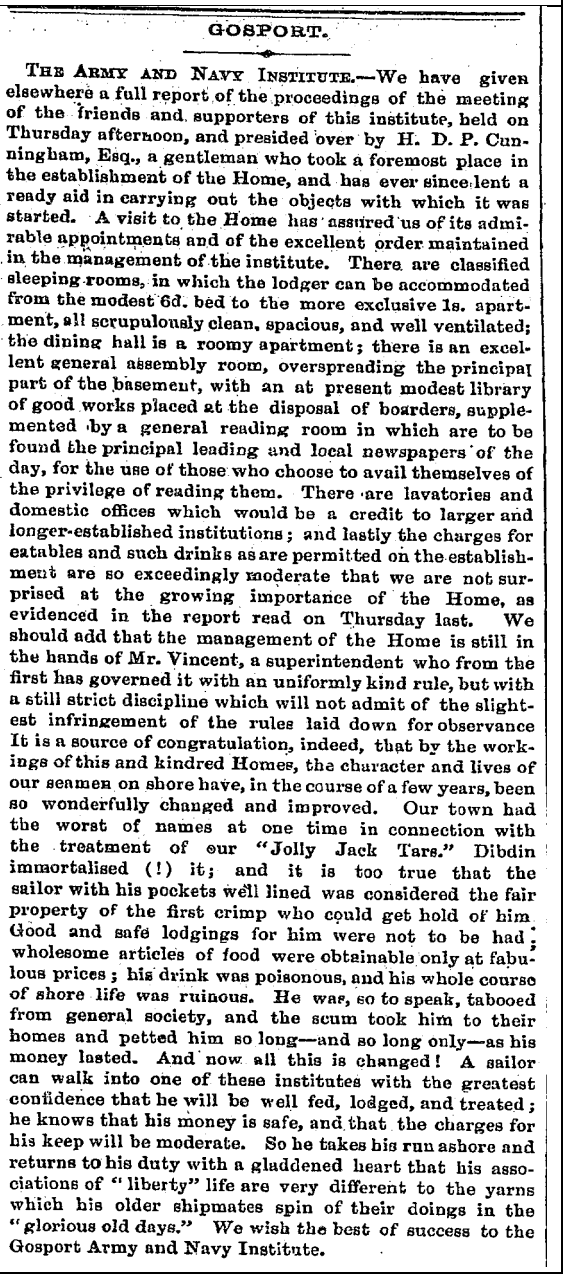
method is capable of. This realisation has epistemological implications, as it places more onus on the researcher to verify that a modelled topic is indeed what it appears to be. OCR quality is directly related to the year in which the text was digitised; more modern OCR software produces significantly better results than what was used when this corpus was digitised in 2003.[30] When available, my advice to researchers would be to always use the most recent OCR transcription.

Topic Viewer Design

In response to the problems posed by OCR quality, during the project, there existed a constant movement towards more articles shown for each topic to get a better understanding of what the topic actually meant. However, when going beyond approximately 20 texts shown, the output format of plaintext became problematic for this purpose, as it became unclear which topic a text was in. In future projects, the way in which the results of the modelling step are analysed needs to be considered earlier in the design process, so a better format can be implemented. Ideally, this would take the form that includes the original image of the article, so bad OCR transcriptions can be circumvented and the researcher has access to the most original edition of the source available. It also keeps the topic

[30] Rose Holley, 'How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs', *D-Lib Magazine*, 15.3/4 (2009) <https://doi.org/10.1045/march2009-holley>; Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', *D-Lib Magazine*, 15.7/8 (2009) <https://doi.org/10.1045/july2009-munoz>; Chirag Patel, Atul Patel, and Dharmendra Patel, 'Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study', *International Journal of Computer Applications*, 55.10 (2012); Kimmo Kettunen and Tuula Pääkkönen, 'Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means', 2016; Mika Koistinen, Kimmo Kettunen, and Tuula Pääkkönen, 'Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing', in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 2017, pp. 277–283; van Wissen, p. 111.

information available while the researcher looks at the texts. An example design based on the experiences of this project is offered in Figure 3.5 below.

<table>
<tr><td colspan="2">Topic 1</td><td>perc of corpus: 40.429082</td></tr>
<tr><td colspan="3">0.003*"bill" + 0.003*"church" + 0.002*"(hear," + 0.002*"irish" + 0.002*"thought" + 0.002*"hon." + 0.002*"hear.)" + 0.002*"law" + 0.002*"earl" + 0.002*"proposed" + 0.002*"how" + 0.002*"because"</td></tr>
<tr><td>Text 2</td><td>article score: 0.998941</td><td>ID: SDLN-1850-04-18-0001-003</td></tr>
<tr><td colspan="2">GOSPORT – THE ARMY AND NAVY INSTITUTE – We have given elsewhere a full report of the proceedings of the meeting ob the friends and supporters of this institute, held on Thursday afternoon, and presided over by H. D. P. Cun- ningham Esp. a gentleman who took a foremost place in the establishment of the Home and has ever since lent a ready aid in carrying out the objects with which it was started. A visit to the Home has assused us of its ami- rable appointments and of the excellent order maintained in the management of the institute. There are no classified sleeping rooms in which the lodger can be accommodated from the modest 6d. bed to the more exclusive 1s. apart- ment, all scrupulously clean spacious and and well ventilated; the dining hall is is a roomy apartment; there is general assembly room in which are to be found the principal part of the basement, which at present modest library of good works placed at the disposal of boarders, supple- mented by a general reading room in which in which are to be found the principal leading and local newspapers of the day, for the use of those who avail themselves of the priviledge of reading them. There are lavatories and domestic offices which would be a credit to larger and longer-established institutions; and lastly the charges for …</td><td> 
GOSPORT.

THE ARMY AND NAVY INSTITUTE.—We have given elsewhere a full report of the proceedings of the meeting of the friends and supporters of this institute, held on Thursday afternoon, and presided over by H. D. P. Cunningham, Esq., a gentleman who took a foremost place in the establishment of the Home, and has ever since lent a ready aid in carrying out the objects with which it was started. A visit to the Home has assured us of its admirable appointments and of the excellent order maintained in the management of the institute. There are classified sleeping rooms, in which the lodger can be accommodated from the modest 6d. bed to the more exclusive 1s. apartment, all scrupulously clean, spacious, and well ventilated; the dining hall is a roomy apartment; there is an excellent general assembly room, overspreading the principal part of the basement, with an at present modest library of good works placed at the disposal of boarders, supplemented by a general reading room in which are to be found the principal leading and local newspapers of the day, for the use of those who choose to avail themselves of the privilege of reading them. There are lavatories and domestic offices which would be a credit to larger and longer-established institutions; and lastly the charges for eatables and such drinks as are permitted on the establishment are so exceedingly moderate that we are not surprised at the growing importance of the Home, as evidenced in the report read on Thursday last. We should add that the management of the Home is still in the hands of Mr. Vincent, a superintendent who from the first has governed it with an uniformly kind rule, but with a still strict discipline which will not admit of the slightest infringement of the rules laid down for observance It is a source of congratulation, indeed, that by the workings of this and kindred Homes, the character and lives of our seamen on shore have, in the course of a few years, been so wonderfully changed and improved. Our town had the worst of names at one time in connection with the treatment of our "Jolly Jack Tars." Dibdin immortalised (!) it; and it is too true that the sailor with his pockets well lined was considered the fair property of the first crimp who could get hold of him. Good and safe lodgings for him were not to be had; wholesome articles of food were obtainable only at fabulous prices; his drink was poisonous, and his whole course of shore life was ruinous. He was, so to speak, tabooed from general society, and the scum took him to their homes and petted him so long—and so long only—as his money lasted. And now all this is changed! A sailor can walk into one of these institutes with the greatest confidence that he will be well fed, lodged, and treated; he knows that his money is safe, and that the charges for his keep will be moderate. So he takes his run ashore and returns to his duty with a gladdened heart that his associations of "liberty" life are very different to the yarns which his older shipmates spin of their doings in the "glorious old days." We wish the best of success to the Gosport Army and Navy Institute.</td></tr>
</table>

Figure 3.5 Mockup of the proposed model for analysis of texts within topic models based on experience. It includes all the features of the improved output in figure 3.4, but includes the image from which the text was digitised, which helps with reading OCR errors. It also keeps the topic information at the top from going out of screen when the researcher scrolls down to discover more texts.

This thesis found that analysing documents this way, with a higher focus on the original text rather than the indicative tokens, was more in line with the historical method than the visualisation-and-analysis workflow that is more commonly used in topic modelling. This is also why this project chose not to use common topic visualisations such as pyLDAvis, as they don't connect the topic model back to the original text, but instead mask it from view entirely. If this project wishes to produce answers to historical questions, and produce a method that may be used by other historians, it needs to stay as close as possible to established historical methods.

Computational Requirements and Implications

The most significant aspect of my experience with the topic modelling tool is related to one aspect: computational speed and the associated memory requirements. These set the boundaries of what this project can achieve in both its methodological and historical research; as part of the goal of this project was to establish these boundaries for a single researcher, it is fitting that these are discussed. This section will cover three main implications of computing power: the speed of querying and article retrieval; the speed and capacity for modelling; and the limitations towards improving the corpus OCR transcription.

The question of access to quick and powerful computers for historical research projects is not new. As Stephen Ramsay observed, humanities departments are not overly endowed with high-performance computing equipment, so often there are practical constraints that research projects have to work within. While this is likely to change over the years as the Digital Humanities

take off further, and as Humanities departments hoover up the cast-offs from the computer science labs, these are still issues that we have to contend with.[31] The initial indexing of the xml files for the elasticsearch engine was done on a server belonging to the computer science department, which could only be used for a limited time. Thus, by the time the script for querying large amounts of text was finished, the actual server had moved to an organisation-standard desktop.[32] This meant that the retrieval of articles from modelling was very slow; querying for up to 10,000 articles for a single keyword and a single year would regularly take 2 to 3 hours. A single keyword for the entire half a century would thus take 5 to 6 days. This meant that retrieval of data was the single most costly (in terms of time) step in the process, and consequently, that the retrieved articles were extremely valuable.

This has a direct impact on the kinds of questions that are asked, and the approaches to solving them. The time investment involved made the selection of keywords and search parameters very important and very conservative: each keyword needed to produce a topic model that addressed the questions in the case study it was used for in some way, and preferably be usable for multiple case studies. In practical terms, this meant not using all possible markers of imperial places; out of all possible options, only those that would produce a sufficient number of articles to model were chosen. The arrival of more powerful hardware halfway through the project meant that text queries that previously took weeks to complete, could now be run in days, essentially removing that particular bottleneck

[31] Ramsay, 'High Performance Computing for English Majors'.
[32] HP Z220. Processor intel i5-2500 quad-core @ 3.5 GHz; RAM 4 GB DDR3 @ 21 Gb/s.

from the workflow.[33] From this point on, it became practical to run more explorative queries which could shed light on narrower aspects of the material, perhaps even explore a single facet of a case study. As a result, the case studies formed around the aspects and facets of imperial life, rather than around the results of one or two queries. Not only that, it also became possible to query for the page scans as well, which on the Z220 had been too slow as this would involve multiple queries following on from the keyword search. This realisation led directly to the visualisation techniques developed in the next chapter of this thesis.

A similar story applies to the topic modelling step undertaken after the articles were retrieved from storage. Here, the key issue was not so much speed, as there was no point producing models faster than the data to create them from could be retrieved, but the amount of memory on the computer preforming the modelling. During modelling, each document that is read into the program adds another 'row' and 'column' to the model, so the model requires more and more space as it is built. This caused the initial machine on which modelling was run, a Z220 identical to the initial server, to run out of memory when handling more than 20,000 articles. Modelling then shifted to the strongest computer that was available, a mid-to high-end home machine.[34] This enabled models to be run with more articles, and when the Z420 arrived, this was used to run more articles per model still. However, the constraints imposed by the size of modelling did have their impact on the process, as they were instrumental in deciding the size of the subsets and number of years that would be modelled at a time. The details of the

[33] HP Z420. Processor intel E5-2680 octuple-core @ 3.5 GHz; RAM 64 GB DDR3 @ 52 Gb/s.
[34] HP Pavillion 570. Processor i5-7400 @ 3.0 GHz; RAM 16 GB DDR4 .

differences between these three machines are shown in Figure 3.6. Had the Z420 been available from the start, different choices could have been made – for example, models would have covered the full half-century instead of the decades they do now. The experiences of this thesis illustrate the importance of considering the hardware constraints in a digital humanities research project as integral to the process, and that they require constant evaluation in light of the research question. Just as a historian reframes their research questions based on the shifting constraints imposed by archival access, so should the digital humanist consider their hardware.

Model	**Processor**	**Memory**	**Upper limit**	**Time taken (10.000)**
Z220	i5-2500	4 GB	20,000	92 min
Pavillion 5	i5-7400	16 GB	120,000	55 min
Z420	E5-2680	64 GB	None Found	7 min

Figure 3.6 Performance of different hardware used for topic modelling. The significant increase in computing power meant a significantly larger number of articles could be retrieved and modelled in the same or less time. This in turn made exploratory queries viable.

Additionally, the way this project used its topic models was significantly shaped by the material it tried to model. At several times in the process, attempts were made to improve the granularity of the topic models and to build additional functionality into the tool, such as better tokenisersers and stemmers. These attempts always ran into the same issue: insufficient OCR quality. The opportunities offered by stemmers and de-hyphenators could therefore not truly be explored, as more often than not, they had negligible effect on the texts being fed into the topic modelling step. As an experiment, an attempt was made to use a spell corrector before applying the stemmer and tokeniser. This was done on a single test corpus of 500 randomly selected articles from the entire archive. Unfortunately, the results proved unsatisfactory for two reasons. First, the spell

correctors were optimised to catch spelling mistakes made by humans on a keyboard, not the kinds of mistakes made by OCR transcription, which has a different lexical pattern. For example, a human will commonly mis-type 'and' as 'abd', while a OCR package is more likely to misspell it to 'ahd' or 'aiid', depending on the typeface. But more importantly, spell correctors provided very little benefit for the amount of computational time they required. A model with spell-checking active takes five times as long to process, for negligible benefits to model quality. A more powerful computer and further optimisation may have made spell-correcting the ingested articles worth it, but it was decided that trying to improve the OCR was beyond the scope of this project.

Overall, the evaluation of the LDA tool presented in this chapter is positive. It preforms as can be expected from a tool produced by a single researcher, even though some aspects, such as the presentation of the resulting model, can still be improved. However, there are some important observations that can help us better understand the tool. First, the data quality is extremely important and significantly shapes the topic models. It imposes a hard limit on how accurate and specific the topic models can be. Second, the computational times involved also influence the models. Not only do they potentially limit the work that can be done to improve the data quality, but they also determine the kinds of models that are generated, by shaping which questions it is possible to ask of the data.

Conclusion

First and foremost, this chapter has shown that it is possible for a single researcher to build their own tools for topic modelling. Doing so facilitates the development of a deep and intimate knowledge of the operation of the underlying algorithm, and allows the researcher to tailor the tool to both their specific inputs and their desired outputs. Yet while it was successful, it also encountered several limits that materialised as a result of conducting research in this way. The main limiting factor it encountered was related to the fact that self-written tools are inevitably less well optimised and less efficient in the use of resources than premade ones; even a single inefficient or suboptimal line of code will slow the entire process down to that level.[35] This has significant implications for the amount of computational resources needed, especially if the tool is intended to handle data at a large scale. These computational constraints have been shown to influence the kinds of research questions that can be asked of a tool, and that access to different resources can greatly change this computational context, forcing a reassessment of the research question(s) that a project can ask.

This chapter has shown the way in which this project implemented its choice of topic modelling algorithm, LDA, and how it tailored it to the specific sources it wishes to work on. It discussed the pre-processing step of transforming the text according to the Bag-of-Words and Tf-IDF models, which privilege rare words over common ones. Crucially, it reported on the modelling settings used for

[35] Donald E. Knuth, 'Big Omicron and Big Omega and Big Theta', *ACM SIGACT News*, 8.2 (1976), 18–24 <https://doi.org/10.1145/1008328.1008329>; Ian Chivers and Jane Sleightholme, 'An Introduction to Algorithms and the Big O Notation', in *Introduction to Programming with Fortran: With Coverage of Fortran 90, 95, 2003, 2008 and 77*, ed. by Ian Chivers and Jane Sleightholme (Cham: Springer International Publishing, 2015), pp. 359–64 <https://doi.org/10.1007/978-3-319-17701-4_23>.

the generation of the topic models, which are the most influential settings shaping the resultant topics. On this dataset, which has severe quality issues, we report a comparatively low number of 20 topics, with a loss of precision in the resulting model, producing better results than the 100- or 200 topic models which previous studies recommend.

This project has justified its decision not to rely on available topic viewers, as those available did not meet the requirements. Connecting back with the article text after the topic model has been generated, instead of working off generated word-topic lists was argued to be a core requirement for using topic models in a historically sound way. This requirement is the direct result of my belief that the best approach to analysing topic models in search of historical evidence is to return as close as possible to the historical method – and thus to the original sources. The advantage of this approach is that even a corpus that would otherwise be too rough to successfully topic model can still produce an average of 80.8% meaningful topics. This finding could substantially change the way topic models are used in historical research.

However, the issues of data quality and topic size have forced the conclusion that the results of topic model analysis have to be treated with a substantial amount of epistemological scepticism. When a model is produced, the text has been transformed so many times, and so many interpretive layers have been placed between the researcher and the source, that topic models should never be taken as conclusive proof of historical reality. As tools, they are well suited to high-level and general exploration of a corpus, and can in that role indicate

interesting avenues of research. They can answer questions about the general context in which a word or set of words were used by giving an indication of different vocabularies with which they co-occurred, and in this role topic models can gesture towards solutions to concrete research questions. However, LDA topic models, as used by this project, cannot provide an objective truth about the past.

This project has found two avenues for future work. The first relates to the quality of the data in the archive. As recent work on other datasets shows, higher quality transcriptions improve the quality of what digital tools in general, and topic modelling in particular, can do on that dataset. The transcriptions that are in this archive were generated almost two decades ago, and while OCR technology has improved significantly, the archive has never been reprocessed. A higher quality initial corpus would make a significant impact on the resulting models, and open the door towards using more advanced linguistic analysis tools in conjunction with topic modelling. The second relates to the possibility to train a topic model to recognise certain linguistical patterns and extract documents that are similar. We could then take a known topic, for example Joke Columns, generated out of a topic model of several hundred curated jokes, and request LDA find us all other such texts in the archive. As such a system is highly dependent on textual nuance having access to an accurate transcription is crucial, but it would open exciting new possibilities for locating irregularly appearing genres which can be hard to find using keyword searches.

Chapter 4: Visualisation

The topic models discussed in the previous chapter have one fundamental epistemological downside: they only analyse textual data. While text is a key part of understanding the meaning of newspaper articles, it is not the only carrier of meaning involved in the reading process. The placement of an article within a paper can also carry meaning. This chapter will present the methodological and theoretical basis for the visualisation of spatial patterns within newspapers, the methods this thesis has developed for visualising the location in which a subset of articles feature in a newspaper, and the process for interpreting the resulting heatmaps. In its first section, it will demonstrate the validity of researching the non-textual aspects of newspapers and show how previous methods to do so have informed this thesis' praxis. In the next section, it will present the implementation of article placement mapping adopted by this project and discuss the design decision involved. Finally, it will argue that the visualisation of article placement allows for the exploration of historic trends in article placement design and the study of article layout, as well as allowing the spatial context of texts to re-emerge.

Method and Theory

To start, let us consider the placement of an article in a newspaper, and why it is worthwhile to investigate. Why was this particular thing put in this particular place on this particular page? It is not a random act, as anyone who has ever opened a newspaper will be able to attest, but the result of a series of conscious decisions made by the editor and printer. The way articles are grouped together on a page, and the way they claim certain pages as their own, imparts a structure and logic on them which contextualises their textual contents.[1] By clustering together news from Paris, Berlin and Vienna, a European news column is formed, in its identity and internal consistency distinct from the cluster of articles discussing affairs in Australia and China. Together, these clusters make up the foreign news section, which might be spread over a page or two towards the middle of the issue. Billig observed that for his sample of modern newspapers, national news tended to precede foreign news, as it is closer to the reader's interest, and articles that are expected to raise particular interest are placed above the fold, or in the top left-hand corner, where our eyes are naturally drawn when we scan a page of text.[2] As journalistic practices evolved over the nineteenth century, newspapers also began to run 'columns', weekly or daily features that always appeared on the same place in the paper, and which became important parts in the paper's identity by their repetition.[3]

[1] Billig, *Banal Nationalism*, pp. 118–19; James Mussell, *The Nineteenth-Century Press in the Digital Age*, Palgrave Studies in the History of the Media, 5 (Basingstoke: Palgrave Macmillan, 2012), pp. 30–31.
[2] Billig, *Banal Nationalism*, p. 117.
[3] James Mussell, *The Nineteenth-Century Press in the Digital Age*, pp. 55, 85.

Understanding the conventions of page and issue layout is crucial for anyone who wishes to understand newspapers. This applies equally to the nineteenth-century editor who needed to sell their product, as it does to the twenty-first century historian who wishes to study their contents. For both, what is written is as important as that which is not. Other historians have argued this importance before, such as James Mussell: "The non-linguistic aspects of any printed object are integral to its meaning. … [They] allow for a publication's identity to emerge through their repetition."[4] Understanding these design aspects is critical to understanding the sources. The placement of articles within an issue, and even within a page, provides the context in which readers consumed and understood these texts. Of course, they can't tell the whole story of the article, but neither can the text. Only in the conjunction of space, form, text and context can we find the full meaning of an article.[5] However, most digital archives hide these aspects by advantaging the text in the way they are designed: searching, retrieving and presenting an article is nearly always done on a textual level.[6] As such, we lose a way to access the context of a page, and are dropped straight onto the (textual) article. This project seeks to address this problem by developing a tool that can reconstruct the spatial context of an article, by allowing the patterns of repetition to re-emerge. It will create a density map of similar articles in the pages of the paper, thus showing not only the repetition of articles but also those that are divergent in placement.

[4] James Mussell, *The Nineteenth-Century Press in the Digital Age*, p. 72.
[5] James Mussell, *The Nineteenth-Century Press in the Digital Age*, pp. 30–32.
[6] James Mussell, *The Nineteenth-Century Press in the Digital Age*, pp. 58–60.

The first step in producing such a tool or method needs to be an exploration of the ways in which the spatial nature of newspaper articles has been expressed in the past. The field of journalism studies in particular has a long history of measuring the placement and amount of space devoted to certain news items and topics. The earliest methods of such quantitative content analysis were developed shortly after the Second World War as a way to deal with the volume of material produced in a mass media context, by expressing it as a spatial measurement: the column inch.[7] Conceptually column inches were nothing new in the business of newspapers, and editors had been using them (and other indicators of space) as ways to gauge the amount of room a piece of writing would use long before the measurement was theorised; advertisements were, for example, priced by the space they took up.[8] But the academic breakthrough of the column inch came in 1949, when D.M. White conducted an in-depth study of the practices of the newsroom editor of an American newspaper, theorising the gatekeeper-role of the hypothetical 'Mr. Gates', and exploring the choices he made in the selection of content. This study sparked a wide range of similar studies.[9]

The use of the 'column inch' or any other measure of space on a page taken up by an article has gone out of favour, though it is still taught as a basic content

[7] Christian Kolmer, 'Methods of Journalism Research - Content Analysis', in *Global Journalism Research: Theories, Methods, Findings, Future*, ed. by Martin Lèoffelholz and David Weaver (Oxford: Blackwell, 2007), pp. 117–30 (pp. 117–19).

[8] Alison Hedley, 'Advertisements, Hyper-Reading, and Fin de Siècle Consumer Culture in the Illustrated London News and the Graphic', *Victorian Periodicals Review*, 51.1 (2018), 138–67 (pp. 140–41).

[9] Paul B. Snider, '"Mr. Gates" Revisited: A 1966 Version of the 1949 Case Study', *Journalism Quarterly*, 44.3 (1967), 419–27 <https://doi.org/10.1177/107769906704400301>; Glen L. Bleske, 'Ms. Gates Takes over: An Updated Version of a 1949 Case Study', *Newspaper Research Journal*, 12.4 (1991), 88–97 <https://doi.org/10.1177/073953299101200409>; David Manning White, 'The "Gate Keeper": A Case Study in the Selection of News', *Journalism Quarterly*, 27.4 (1950), 383–90 <https://doi.org/10.1177/107769905002700403>.

analysis method.[10] While it is hard to ascribe a definitive reason for the ebbs and flows in the popularity of any given research method, there are some indicators as to why these analyses are becoming fairly rare. The most likely explanation lies in the change in access to newspaper archives. When material became primarily accessible through microfilm, the difficulty in measuring column inches increased. Not only did the researcher have to find a way to measure either the film or the area projected on screen, it also meant that article 'size' became dependant on the distance between paper and camera when the image was taken. An inch on screen in one paper was no longer necessarily the same inch as the one in another paper. These changes meant that column inch measurements lost their main appeal: that they were easy, quick, and universal. The proposed solution to this problem, the Basic Space Unit (BSU) also suffered from similar drawbacks.[11] This measure was designed to take into account typeface and size. It could describe a text of a given length, with a given typeface, interline, and point, as a number of Basic Space Units. However, this solution was too convoluted, and tied the measurement too closely to text.

Column inch measurement, then, developed as a way to analyse newspaper content while taking into account its non-textual features. It provided a way to compare the relative amount of space taken up by certain materials, both textual and visual. However, it also failed to be more than a mechanism for comparison between different papers or different kinds of content. A column inch analysis

[10] Janet Woollacott and others, *Mass Communication and Society* (London: Edward Arnold, 1977) <http://capitadiscovery.co.uk/edgehill/items/69710> [accessed 2 February 2018]; Kolmer, p. 124.
[11] Wayne A. Danielson and James J. Mullen, 'A Basic Space Unit for Newspaper Content Analysis', *Journalism Quarterly*, 42.1 (1965), 108–10 <https://doi.org/10.1177/107769906504200114>.

does not explore the space an article covers, but only uses the volume of that space as proxy for importance. For a method that retains elements of the 'visual language' of the original, and that does justice to the visual literacy that the creator would have expected of their readers, other approaches are needed.[12]

Any such approach also has to consider the challenges posed by results of the transformation of newspapers into online digital collections, such as the database used by this thesis. With there no longer being a physical object to take measurements from, the raison d'être for a content analysis based on the spatiality of the newspaper becomes questionable, as a measurement is no longer straightforward. With access to archives generally being conducted through a text-based search process, it became more practical to count the mentions of a term than to measure the amount of space the article mentioning it took up.[13] This new predominance of the text is in its own right indicative of the loss of physicality that is associated with digitised sources, and illustrates how easy it is to lose sight of the actual object. This is not a new point to make, but it bears repeating nonetheless.[14] Illustrating this desire for archival uniformity over material complexity is the humble page number which, in digital archives, is often inconsistent with the actual page number on the digitised image.[15] The problem, then, is to make a migration from the abstracted digital form of the page, which can only ever be a facsimile of

[12] R. W. Burniske, *Literacy in the Digital Age* (Thousand Oaks, CA: Corwin Press, 2008), p. 102.
[13] For an Early example see: Bob Nicholson, 'Counting Culture; or, How to Read Victorian Newspapers from a Distance', *Journal of Victorian Culture*, 17.2 (2012), 238–46 <https://doi.org/10.1080/13555502.2012.683331>.
[14] Charles Jeurgens, 'The Scent of the Digital Archive: Dilemmas with Archive Digitisation', *BMGN-Low Countries Historical Review*, 128.4 (2013), 30–54; James Mussell, *The Nineteenth-Century Press in the Digital Age*, p. 192.
[15] James Mussell, *The Nineteenth-Century Press in the Digital Age*, pp. 82–83.

the physical object, to a measurement of that physical object. The easiest way this could happen is if the archive included in its descriptive metadata a pair of fields for page height and page width, expressed in millimetres. However, this data was never recorded, highlighting the fundamental problem with archival resources: unless the designer, a priori, foresees a way for the archive to be used, it is not coded in. This problem also highlights the power of the curator of the data, and the way that the choices they make about the treatment of the material influence the user.[16] So, in order to make this transition from image to measurement, we have to make clever use of the data that is available. This has to be data that exists in the dataset and which can be accessed at an article level, as the data gathering takes place after keyword selection of the articles we wish to visualise. Relying on such a collage of data requires a tool that is tailored to this particular archive. However, such a tool must also be informed by the work others have done on the same problem, digital or not.

Using an approach akin to close reading rather than a quantitative measurement, there has been some research into the layout and design decisions of nineteenth-century editors. Important work was done by Kevin G. Barnhurst and John Nerone, who theorised the exploration of the 'form' of a newspaper as a historical aspect that had cultural connotations. By interrogating the form of a paper, it was possible to reveal these connotations – and explore their change over time. Their applications of this method focussed on American print culture between the 1880s and 1980s. They showed that contrary to established

[16] For a discussion on the role of the Editor or Curator see: James Mussell, *The Nineteenth-Century Press in the Digital Age*, pp. 127–34.

scholarship at the time, such as work by Garcia and Hutt, it was not technology that drove the visual change of the front page, but the arrival of Modernism as an artistic style, as well as the professionalisation of the newspaper industry.[17] However, their analyses relied on a visual close reading of their selection of newspapers, and sought to reach a general context true for all papers, rather than a specific context of one article.

For non-English language newspapers, the little work that has been done has predominantly focussed on Scandinavian and Baltic newspapers; with secondary literature written in those languages. Thus, due to its most defining texts being in Finnish or Russian, this scholarship has found little international traction, and can only be accessed through reflections on these texts by other scholars who write in English and can also access these texts. One such gateway scholar is R. Kurvits, who has worked on the visual form of Estonian newspapers in comparison with Russia and the other Baltic states. In her work, she applies the Appearance Spiral Model developed by P. Mervola to Estonian newspapers to explore the cultural and economic context in which they were produced and consumed. This model argues that the form of the newspaper is not directly informed by the economic, social, and cultural environment in which it is read, but that the volume of content serves as a mediating factor between the cause and the visual result.[18] Kurvits showed that the Estonian Press underwent visual change as

[17] Allen Hutt, *The Changing Newspaper: Typographic Trends in Britain and America 1622-1972* (Gordon Fraser, 1973); Mario R. García, *Contemporary Newspaper Design: A Structural Approach* (Prentice-Hall, 1987); Kevin G. Barnhurst and John Nerone, *The Form of News: A History* (New York: Guilford Press, 2002), pp. 188–92; Kevin G. Barnhurst and John C. Nerone, 'Design Trends in US Front Pages, 1885–1985', *Journalism Quarterly*, 68.4 (1991), 796–804 (pp. 802–3).

[18] Pekka Mervola, *Kirja, Kirjavampi, Sanomalehti: Ulkoasukierre Ja Suomalaisten Sanomalehtien Ulkoasu 1771-1994* (Suomen historiallinen seura, 1995), I; Roosmarii Kurvits, 'The Visual Form of Estonian Newspapers from

a result of increasing Russian (and Soviet) influence on the country during the nineteenth century, but that this evolution was mediated through the variations in volume of certain types of newspaper content, mainly advertising.[19] However, all these means of analysis still hold two downsides. First, they are textual in their own way, describing the visual form into words before analysing it, and returning a result that is also textual. Second, the methods used by these scholars are akin to close reading of the form, which does not scale well with the amount of articles that this thesis hopes to analyse. Kurvits, for example, manually measured the area of the physical newspapers, and was thus limited to three samples per year for her 45-year study.[20] In order to avoid having to manually measure newspapers, this thesis developed a computational approach using the visualisation of newspaper spaces.

In his 1993 book, F. Haskell warns that the analysis of images too often still relies on text. Visual works are transcribed into texts and analysed as texts, because text is the standard method of communication in academia.[21] While Haskell was writing in the context of art history, this thesis still considers his warning applicable and wishes to analyse this spatial and non-textual element in a spatial and non-textual way. Otherwise, it would not be the space that the tool we developed helped

1806 to 1940 and the Appearance Spiral Model', *Nordicom Review*, 29.2 (2008), 335–52 (p. 336) <https://doi.org/10.1515/nor-2017-0195>.

[19] Roosmarii Kurvits, 'The Visual Form of Newspapers as a Guide for Information Consumption', in *Things in Culture, Culture in Things*, ed. by A. Kannike and P. Laviolette, Approaches to Culture Theory, 3 (Tartu: University of Tartu Press, 2013), pp. 172–203.

[20] Kurvits, 'The Visual Form of Estonian Newspapers from 1806 to 1940 and the Appearance Spiral Model', pp. 337–38.

[21] Francis Haskell, *History and Its Images: Art and the Interpretation of the Past* (New Haven, CT: Yale University Press, 1993), p. 3.

understand, but a (descriptive) textual representation of that space. Therefore, the end result of the analysis step has to be a visualisation.

Both History and the Digital Humanities have a long history of producing visualisations to facilitate the generation of new knowledge. Historians often point towards Minard's graphic of Napoleon's journey through Russia as the instigator of graphical presentation of a wide range of information for the purposes of producing an understanding of past events.[22] In his work, visualisation served as the endpoint of research, as a way to present findings. Later, historians in the nineteenth century employed visualisation techniques to great effect, but by the early twentieth century they fell out of favour, especially amongst the quantitative historians that had employed them. "Numbers, parameter estimates, and, especially, standard errors were precise. Pictures were—well, just pictures: pretty or evocative, perhaps, but incapable of stating a "fact" to three or more decimals."[23] However, the arrival of personal computers and digitised data enabled a new wave of scholarship to utilise visualisation in ways that were previously impossible.[24] This became part of the Digital Humanities quickly, and the field has become a nexus of research in humanities visualisation.[25] However, many of these

[22] Charles-Joseph Minard, 'Carte Figurative des Pertes Sucessives en Hommes de l'Armee Francaise dans la Campagne de Russie 1812 - 1813' (Paris: Regnier et Dourdet, 1844), Bibliothèque numérique patrimoniale des ponts et chaussées, Ecole nationale des ponts et chaussées, Fol.10975; Michael Friendly, 'Visions and Re-Visions of Charles Joseph Minard', *Journal of Educational and Behavioral Statistics*, 27.1 (2002), 31–51 <https://doi.org/10.3102/10769986027001031>; Michael Friendly, 'A Brief History of Data Visualization', in *Handbook of Data Visualization*, by Chun-houh Chen, Wolfgang Härdle, and Antony Unwin (Berlin, Heidelberg: Springer Berlin Heidelberg, 2008), pp. 15–56 <https://doi.org/10.1007/978-3-540-33037-0_2>.

[23] Michael Friendly, 'Milestones in the History of Data Visualization: A Case Study in Statistical Historiography', in *Classification — the Ubiquitous Challenge*, ed. by Claus Weihs and Wolfgang Gaul (Berlin/Heidelberg: Springer-Verlag, 2005), pp. 34–52 (p. 6) <https://doi.org/10.1007/3-540-28084-7_4>.

[24] Friendly, 'Milestones in the History of Data Visualization', p. 7.

[25] M. Jessop, 'Digital Visualization as a Scholarly Activity', *Literary and Linguistic Computing*, 23.3 (2008), 281–93 <https://doi.org/10.1093/llc/fqn016>; Stéfan Sinclair, Stan Ruecker, and Milena Radzikowska,

visualisations still rely in some form on textual data, which, as will be discussed below, can be problematic.[26]

The visualisations developed by this project bear some resemblance to the work of M.H. Beals in her article on the re-use of newspaper text in colonial newspapers. She developed a visualisation of newspaper columns to show the types of articles that were placed at various places within five specific issues of the *Caledonian Mercury*, while hunting for items that were copied from foreign papers either with or without attribution (Figure 4.1). As she lacked data on the position of an article on a page and could only retrieve the digitised text and the sequence in which it had appeared on the page, she had to infer the article length and placement in columns from the article word counts, relying on the average word count of a column for its length. Beals also noted that her visualisation did not allow for comparison between papers across different years, and only allows for the discovery of patterns in the broadest categories.[27] But the key issue with Beals' visualisation is that they scale poorly, as all the articles on a page need to be classified to produce the visualisation. Missing even one article will change the number of words in a column, and thus stretch all the content that has been included to fill that column, leading to articles being shifted around. And even if

'Information Visualization for Humanities Scholars', 2013 <https://doi.org/10.1632/lsda.2013.6>; Anne Burdick and others, *Digital_Humanities* (Cambridge, MA: MIT Press, 2016), p. 9; Erik Malcolm Champion, 'Digital Humanities Is Text Heavy, Visualization Light, and Simulation Poor', *Digital Scholarship in the Humanities*, 32.suppl_1 (2017), i25–32 (pp. 25–27) <https://doi.org/10.1093/llc/fqw053>.

[26] Elijah Meeks, 'Is Digital Humanities Too Text-Heavy? | Digital Humanities Specialist', *Digital Humanities Specialist*, 2013 <https://dhs.stanford.edu/spatial-humanities/is-digital-humanities-too-text-heavy/> [accessed 5 February 2020].

[27] M. H. Beals, 'Close Readings of Big Data', pp. 621; 623,629.

all these steps are taken correctly, a researcher still has to look at tens of thousands of abstracted pages in order to draw inference about practices of article placement.

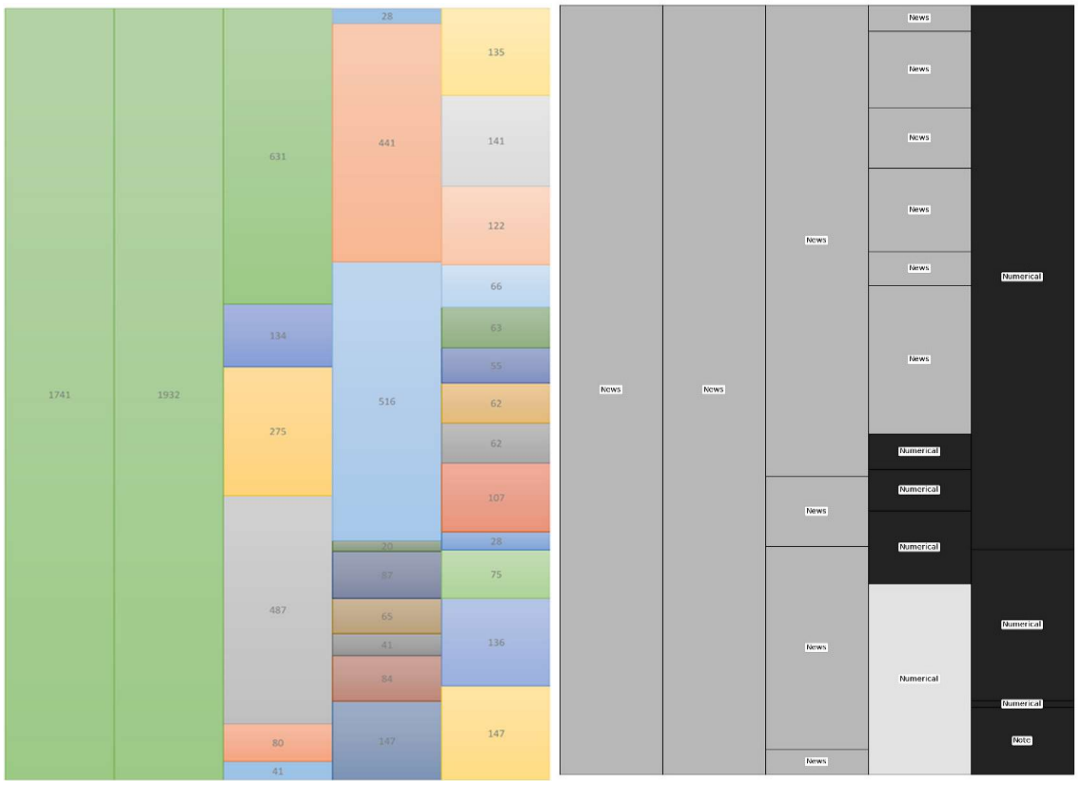


Figure 4.1 Visualisation of two pages of the Caledonian Mercury, with the labels showing the underlying word counts (left) and classification by type and source (right).[28] *Beals' visualisation relies on the word counts of the articles to generate stacked bar graphs (left). These values can then be assigned categories, in this example (right) an indication of where the article came from.*

The section above has shown the importance of the non-textual aspects of newspaper sources. Not only does the placement of the article on the page define the textual and visual context in which it has to be read, it also reveals the meta-textual structures that the editor intended for the text to be understood in. Additionally, through the repetition of these elements, periodicals produced their own identities. Understanding the pattern of repetition allows a better understanding of the periodical as a whole and the article in particular. This

[28] M. H. Beals, 'Anatomy of a Newspaper: The Caledonian Mercury, 20 June 1825', *MHBeals.Com*, 2017 <http://mhbeals.com/anatomy-of-a-newspaper-the-caledonian-mercury-20-june-1825/> [accessed 22 October 2019]; M. H. Beals, 'Close Readings of Big Data', p. 624.

understanding has driven other researchers to develop ways to measure and understand non-textual context, specifically space. The first of these, the Column Inch, is not suitable for the goal of this thesis, as it does not represent the space of the page but the space of the article, which makes it only suitable as a method of comparison of relative article importance. Essentially, column inches use space as a stand-in for text and thus do not truly fulfil the desire for a spatial analysis method. The close visual reading techniques discussed, such as those used by Barnhurst and Nerone, Kurvits, and Beals, while very effective at generating a close understanding of the page as a whole, do not scale well enough to larger datasets for the purpose of this thesis. Thus, the method this thesis develops has to be both scalable and be a real representation of the article in the context of its page.

Architecture

In the above section, this thesis has shown the previous work that has been done on the visualisation of article placement and the spatial nature of newspaper contents. Compared to this previous work, this thesis has one major advantage: it has access to both the textual and the spatial data of each article, which allows it to correlate article and place directly by using the coordinates of the article on the page image. This means that unlike most examples, it does not have to rely on an approximation through word count. Instead, this tool can rely on the real position on the physical page. By taking a tally of the amount of times articles from a corpus occupy any given position, we can arrive at their occurrence density. This information is then visualised in a heatmap, where place on the graph corresponds to place on the page.

Overall, the process of generating the visualisation is divisible into four main steps:

1. The *(Meta)data lookup* collects the data that the visualisation corpus needs from the different parts of the archive with a keyword search.
2. The *Scaling Step* prepares the data for visualisation by harmonising the coordinate systems that each article is in, and resolving, as far as possible, ambiguities in the assignment of blocks of text to multiple pages. The harmonised coordinate systems are then handed off to the *Table-of-Occurrence Generator*.
3. The *Table-of-Occurrence Generation* tallies the number of articles that occupy a certain space on a certain page into a tabular format.
4. The *Heatmap Visualisation* generates the final image.

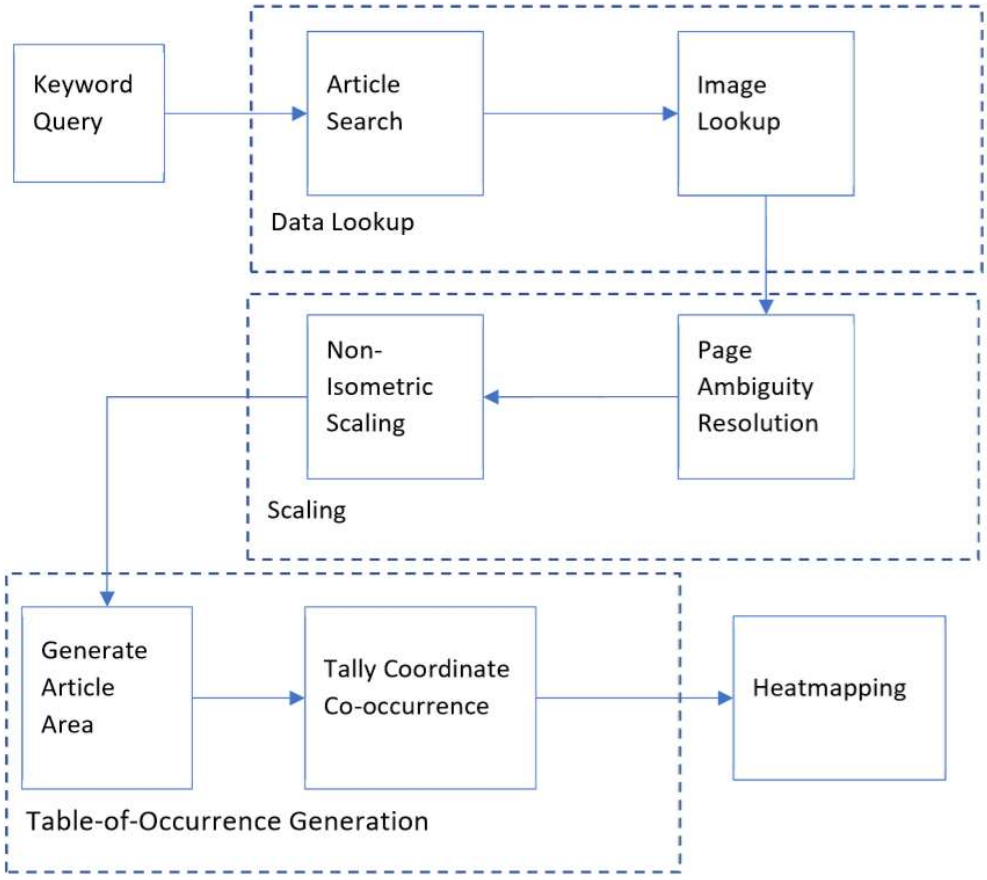


Figure 4.2 Graphic representation of the steps involved in generating article placement heatmaps. Each dashed rectangle corresponds to one of the major subsections of this chapter.

Metadata Lookup

The first step to visualising the placement of newspaper articles is to create a subset out of the corpus. Unlike topic modelling, where the use of subsetting serves both a practical and epistemological goal, in the case of the visualisation it is purely epistemological. If we were to visualise all the articles in the corpus, we would produce a representation of all articles in the corpus, which means there would be no areas of greater intensity from which to draw any conclusions. Thus, the creation of a subset of articles whose positions might produce insight into the

subject of choice, or which might answer a research question is a necessity. For example, if we wanted to visualise patterns in the placement of poetry columns across four decades of a specific newspaper, we would first need to identify and create a subset of all the poetry columns that we wish to measure. Theoretically, this subset could be assembled by manually identifying each column, but in order to visualise long-term trends and deal with large bodies of journalism, we need to identify material for our subset using digital search methods.

The first step in this process, keyword selection, requires intense consideration as to the choice of keyword. The words chosen must cover the subject, but leave as little semantic ambiguity as possible. In some cases, such as the measurement of a consistently titled recurring column, this method can be accomplished with a single keyword or phrase. However, it proved impossible to select single keywords that did justice to the complexity of the research questions that underpin the case studies on expressions of imperial identity used by this thesis. As a result, the tool can gather sources using several keywords and then merge the results, with multiple occurrences of the same article appearing with different keywords only counted once. The specific terms chosen for each experiment will be discussed in the relevant case studies.

The visualisation corpus is first generated by the keyword search, which extracts the article ID's, article coordinates, and the article text.[29] Next, the list of article ID's is used to perform a reverse lookup of the source images, by trimming the article ID's into page ID's. This produces a link to the page's raw image file on

[29] The properties of these attributes may be found in more detail in Chapter 2.

disk, which is then accessed and the image size in pixels and resolution in dpi pulled from its metadata. [30] This process is represented in Figure 4.3

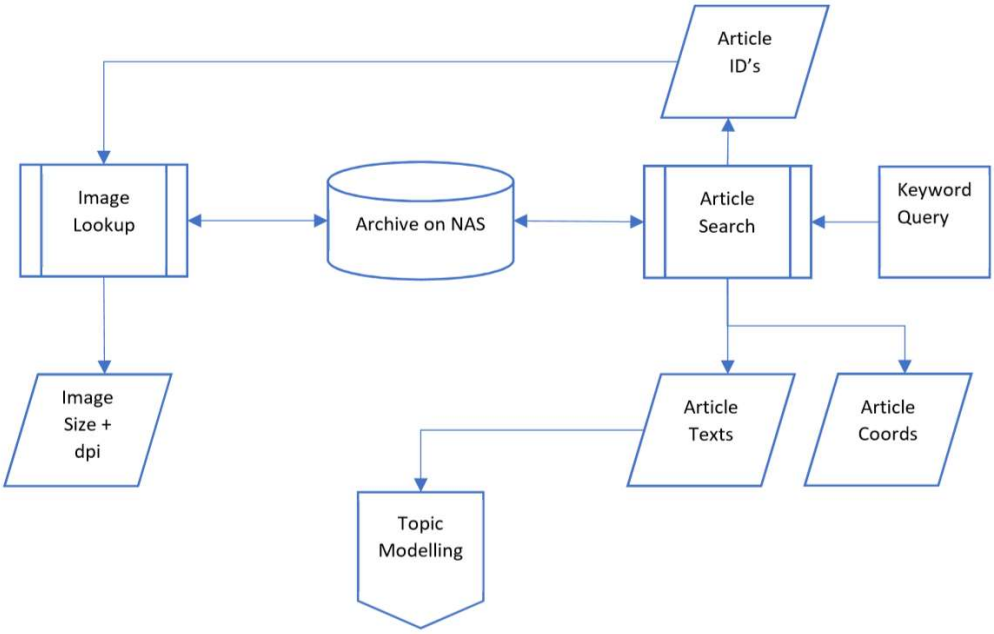


Figure 4.3 Diagram of the retrieval of article- and page data. The keyword query is used to subset articles, similar to the process used for the topic modelling. The article ID's (top) are then retained to preform reverse lookup on the photographs that correspond to the page containing the article.

The keyword search query is executed the same way as for the topic modelling described in the previous chapter: an Elasticsearch query from an external controlling laptop to the NAS over WiFi. The response returns a list of json-records, (see appendix), which it stores in the 'article list'. It simultaneously produces a 'page list'. This is created by taking the article ID and removing the last four characters; as the ID's are structured in the form of <Paper code>_<date>_<page>-<3-digit article number>. Each new batch of articles is referenced to the growing page list, to ensure only the ones that are not yet recorded are added. At the end of the data lookup, we thus have two lists: one with page ID's and one with the combination of article ID and coordinates.

[30] For opening the image files the tool uses the PILLOW package.

Scaling and Harmonisation

At this point, the information we have on an article is simply an ID and an associated set of points that mark out the edges of the rectangle covered by the article. For single-page articles, this is all that is needed here. For articles spanning multiple pages, the next step is to sort out which page each set of points belongs to. This is a particular issue as the coordinate sets do not have 'page' as a component, and therefore there is no way to be sure on which page they belong. However, when we assume that the sets of coordinates are sequential, and that no article spans more than two pages, we can look at the coordinates themselves to determine their page. If the right-most point of the one rectangle is further left i.e., its x-value is smaller, than the leftmost point of the next rectangle, the second must be on the page before the first. These assumptions are, of course, not always true: small-form papers such as the *Pall Mall Gazette* were explicitly not chosen for visualisation because their articles cross over multiple pages. However, it is the best that can be done within the constraints imposed by the archive's designers. With the question of pagination answered, we can begin to address the issues of page size and scaling: for this we need to extract some data from the actual pictures of the page.

The list of pages is used to query the server for the images of these pages; these queries return the location of the image on the server, in a form similar to a web url or a file path. This is then used in a simple web request to gain access to the page-picture. Once the coordinates of the article as well as the dimensions and resolution of the page it is on are known, the process of normalisation can take place, which involves scaling the images and associated coordinate systems to fit

within the same frame of reference. The newspapers contained in this archive are far from uniform in size, ranging from the regular-sized dailies, which were typically 12 ¼ by 18 ¾ inches, to the much smaller *Pall Mall Gazette*, which was only 6 by 9 inches.[31] Additionally, even if the pages were the same, the idiosyncrasy of the digitisation process mean that a page might have gained or lost an inch or two to the binding or simply to the way it was placed on the scanner. Compounding this issue is the fact that the coordinates of an article are in pixels, not in inches, and there is no universal translation between these two units. The reason for the use of pixels is simple: the coordinates were never intended by the archive creators to be used this way. All it was designed to do was to provide a visual indication to the user where on the digitised page the article they were looking at was located. This means these coordinates were always intended to be image-specific, and there was no need for uniformity of any kind between the images.

Both of these issues need to be addressed in order to compare different articles on different pages with one another. First, the frames of reference in which the coordinates of the article exist need to be normalised; that is, both coordinate systems need to be given the same meaning. The normalisation process takes care of one key disparity in the archive: image resolution. The archive itself was constructed in stages, and between the two parts of the archive being digitised, aspects of the digitisation infrastructure were changed, resulting in later parts being digitised on higher-resolution scanners. This means in practice that some pages were scanned in 300 dpi, and the latter part in 400 dpi. Consequently, coordinates

[31] M. H. Beals, 'Close Readings of Big Data', p. 621.

on a page from part two of the archive are 1.33 times larger than those in the first half. Or, in other words, pixel 100,100 on a 300 dpi image equals pixel 133,133 on a 400 dpi equivalent. This provides us with an unworkable situation for visualisation, if the goal is to compare an entire print run of a paper, or even multiple different newspapers for a year or decade. Using the image size and dpi information of the page from which an article comes, we transpose the coordinates of the article onto an 'ideal' newspaper page at a resolution of 400dpi that is the same for all articles, by multiplying the 300 dpi coordinates with 1.33. Next, the result is scaled non-isometrically to fit the ideal page's width and height, as shown in Figure 4.4. A separate X and Y scale are used for this, to assist in columns over different pages aligning with each other. The size of the ideal page was chosen so it has an aspect ratio that is equal to modern A4 for easier printing.

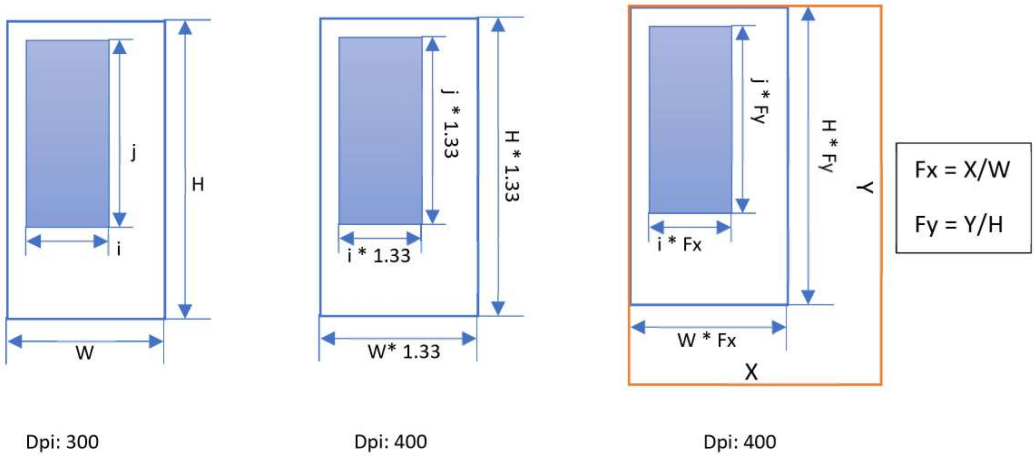


Figure 4.4 Stages of Normalisation and Scaling. Left: Original Page. Centre: Normalised for resolution. Right Non-Isometric Scaling for size.

Table-of-Occurrence Generation

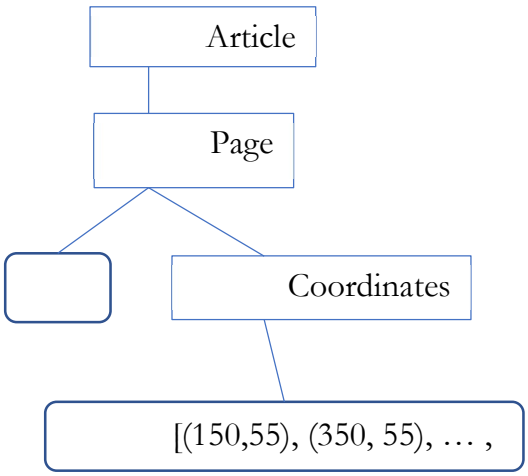
Once the pages have been normalised and scaled, and the coordinate systems harmonised, it remains to generate the table of occurrence from which the heatmaps are drawn. These are the values that inform the intensity of the heatmap,

and represent the amount of articles that occupy any given space. However, we first need to determine the coordinates of each pixel within the article's area. The form in which they are recorded in the archive is as an array, or set of arrays, of four values: left, right, top, and bottom. The four possible combinations of these values represent the corners of the rectangle(s) the article covers, but not the pixels within. These thus have to be calculated in order to fill in the table of occurrences. However, involving every pixel was found to be very impractical: it increases computation time and memory used substantially, but when analysing it offers no additional benefit. A single pixel in our scaled image represents 0.0025 inches square (about the thickness of a human hair) – much too small to notice. The tool therefore uses 200 by 200 pixel blocks in its calculations, as it saves significant resources. These represent an area of 0.5 inch square in physical terms, which is small enough to show the detail we need, but big enough to not squander computer time.

This thesis used a pair of loops to arrive at an array of tuples of all possible x and y coordinates. These have the advantage over other options that the reduction to 200-pixel blocks can happen before all the possible coordinates have been calculated, which avoids investing resources which will go to waste anyway. These, however, are more difficult to code, but once they were operational produced the same output as other methods, but markedly quicker. Supplied below is a pseudocode presentation of the process used.

```
For x_range [left, right]:
      While last_point_added_x < right:
            New_x = last_point_added_x + 200
For y_range [top, bottom]:
      While last_point_added_y < bottom:
      New_y = last_point_added_y + 200
Return (new_x, new_y)
```

At this point in the process, we have generated a collection of articles, with each article containing all the necessary information needed to generate the occurrence tables: a harmonised and unified set of coordinates covering each point within the bounds of an article, with each rectangle assigned to their correct page, for each article in the visualisation corpus. An example instantiation of this form is illustrated in figure 4.5 realised for an article consisting of a single column on page 3, covering the area from point (150,55) to point (950,1255).



X	Y	N	Page
150	55	1	3
350	55	1	3
…	…	…	…
950	1255	1	3

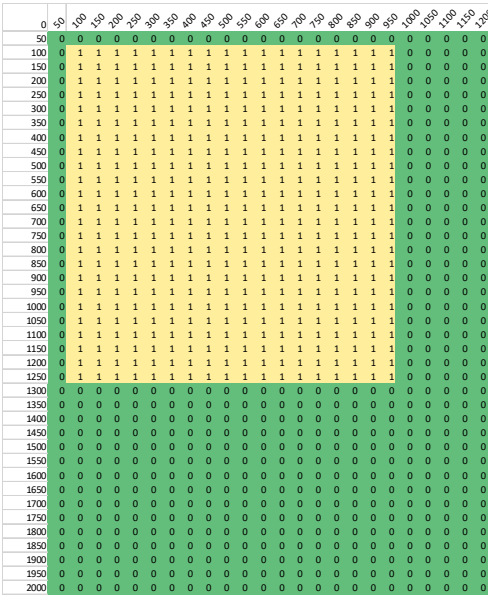


Figure 4.5 Example instance of a single page article stretching from 150,55 to 950,1255 during different stages of the visualisation process. Top left: schematic representation of the data structure. Top right: tabular representation of position data with number of observed articles. Bottom left: heatmap showing underlying number of observances and x- and y-coordinates.

For each article in the collection, the coordinates are passed to the table generation script. It iterates over the collection of articles and adds all their combinations of place and page into the table. If such a combination already exists, it instead increments the value it finds in the field tallying the number of occurrences with 1. Visualising this table using colour intensity tied to table value produces the heatmaps sought by this project.

The form of the occurrence table follows Wickham's convention on tidy data, which is designed to minimise the semantic and analytic uncertainty when dealing with a dataset. It sets out two guidelines for creating such a table: each variable forms a column, and each observation forms a row. [32] Thus, this table has the X and Y coordinate of the 200-pixel square, the amount of times an article occurs in that square, and the page number on which the article sits as their own columns. The program iterates over the articles it needs to visualise, and looks up

[32] Hadley Wickham, 'Tidy Data', *Journal of Statistical Software*, 59.10 (2014), 1–23.

the space occupied by the article in the table. A model of the table is shown as Figure 4.6 below. If an article has already been observed in that place, it increments the number of observations by one; if not, it adds a new row to the table with the article's position.

Row	Contents
X	X-coordinate of the 200 pixel square
Y	Y-coordinate of the 200 pixels square
No. Articles	Number of articles that occur in the square corresponding to these X and Y coordinates
Page	The page from which the article originated, as given in its archival metadata
Topic	Which topic in the topic model an article was assigned to. If assigned to multiple topics, the one with the highest certainty is chosen. Optional.

Figure 4.6 Model of the table underlying article placement visualisation. Following the tidy data convention,

Image Generation

With the data scaled and presented in a tidy format, it only remains to generate the images. For this the thesis relies on Python's matplotlib package, extended by the seaborn package, which has provisions for generating heatmaps. To visualise a single heatmap of a single page, the process is as simple as handing the data to the 'heatmap' function and letting it generate the image. It will take care of the intensity of the scale, with the darkest hue corresponding to the highest value in the data automatically.

A key problem was visualising multiple pages at the same time, while maintaining a uniform density scale between the graphs. If all pages were simply visualised on their own without such a precaution, each would default to a local scale, and the same hue on the heatmap could then indicate widely varying numbers of articles observed. Neither seaborn nor matplotlib were designed to support

generating multiple heatmaps with a common scale, so a workaround using the manual scale settings had to be found. This necessitated discovery of the maximum number of articles observed before generating the heatmaps, in order to place a set value as the maximum, by generating the table of occurrence for all pages in the paper, and using the maximum value from the incidences column. This results in a sequence of heatmaps, which each represent a page, but which share a common colour scale. An example of this is included as figure 4.7.

If the subsetting has been done based on a time range, a final manual step can be used to create a figure which shows a change over time. For this, we generate each row independently with its own internal scale. These are then stitched together in a graphical editing program in the correct sequence. In theory, it would be possible to generate all these graphs simultaneously with a shared scale, by adding a column to the table containing the subset each article belongs to, e.g. 'decade'. However, this was not done due to the required use of resources for such a visualisation.

One of the issues on which the design of the heatmap can draw from literature is its colour selection. While at first glance this may seem to be a minor detail, the selection of colours can significantly alter the way a graph is interpreted.[33] For example, the most commonly used colour scale on a heatmap, the 'rainbow' colour map, is also considered one of the worst, as it distorts the relative intensity of data points.[34] Good data visualisation tells a story about its data

[33] Bernice E. Rogowitz, Lloyd A. Treinish, and Steve Bryson, 'How Not to Lie with Visualization', *Computers in Physics*, 10.3 (1996), 268–273; C.G. Healey, 'Choosing Effective Colours for Data Visualization', in *Proceedings of Seventh Annual IEEE Visualization '96*, 1996, pp. 263–70 <https://doi.org/10.1109/VISUAL.1996.568118>.
[34] Borland and Taylor.

that would otherwise be difficult or even impossible to uncover if it was presented in another form.[35]

Various forms of colouration were experimented with. The initial visualisations used a schema of increasingly dark shades of a single colour. This was highly effective at showing the areas with the strongest presence of articles, but it was found that in places where there were only slight variations in the number of article occurrences, such as on page 2 or page 5 of the visualisation in Figure 4.7, the slight variation in shade was not easily spotted. Thus, the visualisations all use a three-tone colouration scheme of converging colours. The Yellow-Green-Blue scheme that was finally chosen combined ease of interpretation with and aesthetically pleasing form.

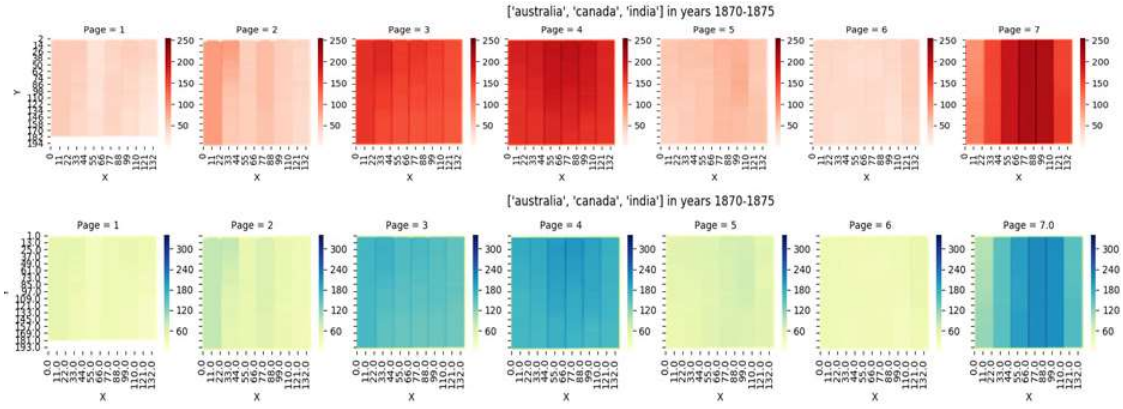


Figure 4.7 Article placement for three 'Imperial' keywords in Reynold's Newspaper for 1870-1875 shown to illustrate interpretative difference in colour. The variance in colour scheme is particularly noticeable for pages where the amounts of articles are close to each other.

Reading these images is complex and requires experience. Each square heatmap represents a page, which is compressed into a unified scale. This is more apparent in the vertical than in the horizontal, as this makes the columns more

[35] Mahmud and others, p. 105.

pronounced. The columns on the page form by themselves, as the article placement data of each overlaps slightly with the adjacent column. This has the effect of creating a darker area delineating the column. A darker hue of blue represents a stronger presence of the subset in that space. In the example above, the strongest concentration is in the bottom-right corner of page 8, with a medium concentration on pages 3 and 4. All pages have some level of article occurrence; if there are no articles (the observance count is zero), the space on the image would be white (for example at the bottom edge of page 1). In practice, these snapshots, be they per year, per decade, or per month, can be stacked on a page, showing the way focal points of keywords move through the title in their repetition.

Topic Modelling and Visualisation

For a final experiment this project sought to combine the visualisation tool with the topic modelling tool. The abstraction of the table of occurrences as a table of tidy data means that the visualisations are easily extended with additional features, such as topic models. This results in a visualisation per page per topic, of which a model is shown as Figure 4.8. These visualisations are much more memory-intensive to run than the regular ones, as they increase the length of the table significantly. However, they are a useful tool for labelling the topics that have been generated, as occasionally the meaning of a topic is also interlinked with the place in which it occurs. For example, during the case studies the visualisations found a column of readers' letters in *Reynold's Newspaper* that would appear as writings on a wide range of subjects from a textual standpoint, and thus their cohesion would be lost in the archive; yet using the visualisation shows they originate from a similar place in a paper, which allows them to be investigated as a group.

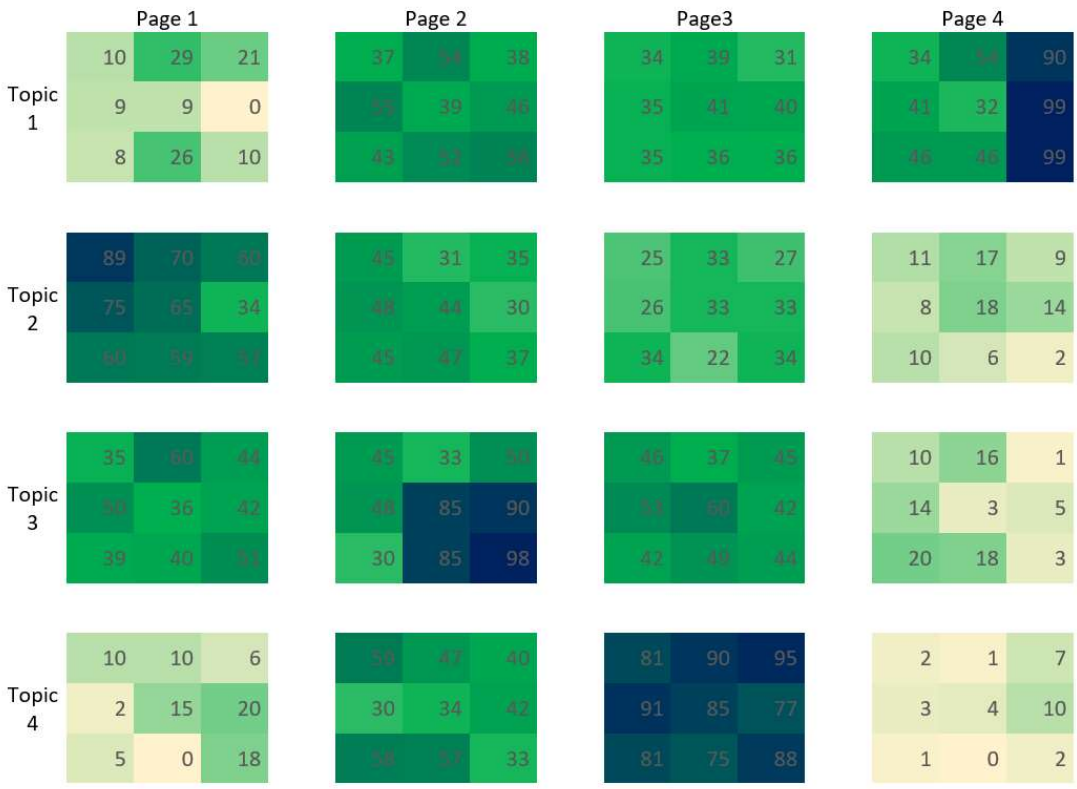


Figure 4.8 Example mock-up of article occurrence visualised per topic for a 4-topic, 4-page publication. Numbers on the image represent number of articles observed. In this example, when trying to understand topic 3, the historian would have to look at the page map for the bottom-right area of page 2, where they might find it contains a joke column.

In the above example, a set of papers of four pages in length, with three columns each, was modelled in four topics. With the information from these visualisations, assigning a meaning to the topic model becomes easier, as it becomes possible to rely on the spatial context of the text, in the same way that the original readers could organise these texts when they read them. For example, Topic 1 concentrates itself in the last column of the last page. The historian can then focus on that column and investigate what that space signifies to this publication; a recurring advert, a joke column or the weekly railway timetable. We can then also read the texts in this topic in that spatial context. For example, the texts in topic 1 are concentrated on the front page, and should thus be read in that context. Of course, the realities of topic modelling mean that cases are rarely this clear-cut, but the same process applies.

Interpretative tool: the column map

Once the figures have been generated and the placement of the articles in the selected subset revealed, we need to interpret the findings. Fundamentally, the question that needs to be answered is this: what does an area of the newspaper represent? Once we know if a large cluster is on the advert pages or the political news column, we can address the historical questions that insight generates. For this purpose, this thesis has developed a supplementary interpretative tool it calls a column map. This map is manually constructed, based on generalised column identification from the close reading, or rather close viewing, of a sample ten papers per five years. While this sample size is small, the maps were created in response to visualisations over whole print runs finding the layout highly consistent over time, barring some clearly recognisable redesigns. Each column in those papers was categorised, and then these were merged into one column map for a certain design cycle in the newspaper's history. If one column always held adverts, it was marked as 'adverts' on the column map, if it was a mixture of political news and economic news, both would be included and the map would note the relative frequency of these two types of content. These maps use the same rough classification system as the topic model topics, so they are interoperable. Using these, the researcher can cross-reference the areas of high intensity and high occurrence of articles in the heatmap with the categories of the column map.

For example, in the column map for *Reynold's Newspaper* in 1860-1875, Figure 4.9 below, each colour represents the presence of a certain kind of content. The front page is left blank, as over the sample, all six different categories were found in that space. The half-coloured columns represent different content in that

column in different issues, with content overflowing into adjacent columns and pages when a significant event takes place. This overflow is particularly notable in the foreign news (red), which overflows to the last columns of the previous page or the first columns of the next page on occasion. When a visualisation produces a concentration of articles in one specific area, the column map can be consulted to get an initial idea what kind of content an area of high or low activity on the heatmap represents in the paper.

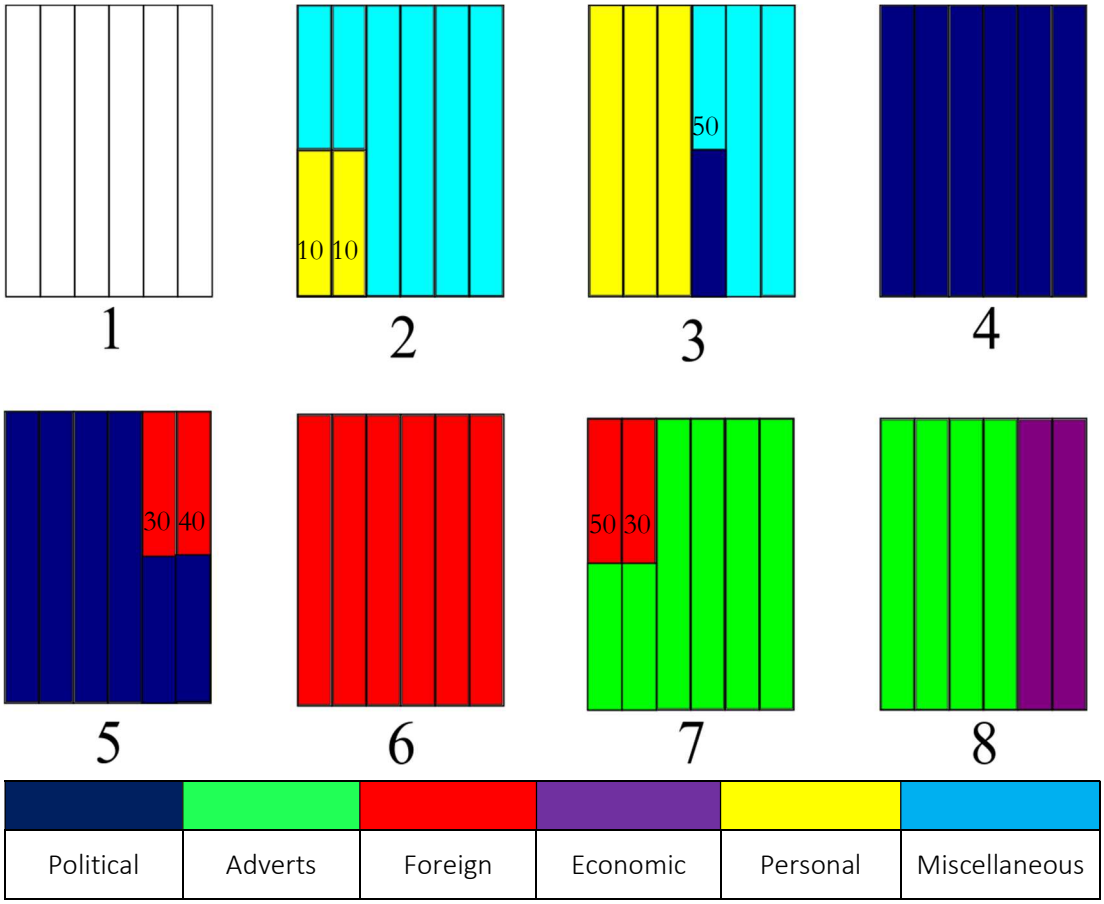


Figure 4.9: Column Map for Reynold's Newspaper between 1885 and 1895. The column map was generated using twenty newspapers as a sample, through manual reading and classification of the content of each column.

These column maps are necessary tools for the interpretation of the heatmaps, but they do mean that there is still a need for human (close) reading of the sources. While they have been verified on a sample of newspapers, it is always possible that these were atypical, or that columns moved around over time.

However, this requirement is orders of magnitude smaller than in existing methods. Compared to the work by Beals, who had to categorise each article on each page she wished to visualise by hand, the column map requires only a sample to be treated in such a way.[36] Thus, while it is not a fully automated system, it still means the problems with scaling Beals' approach to article mapping are reduced. Consequently, we can retain the advantage of a data-driven approach, but the need to use human-generated column maps provides a welcome counterpoint of human control and interpretation.

This section has presented the way in which this thesis visualises article placement and analyses the resulting graphs. Having the option to create these graphs is valuable to researchers. Not only does it provide a way to interact with the source material through a different, non-textual angle, and thus facilitate new insights into the material, it also allows for the reclamation of a little bit of the materiality of the original document. Non-textual aspects, like the design of a page, are the first to disappear from view when the mode of access to the archive is changed.[37] Typically, these new modes of access are linguistic, which limits the way non-linguistic information can be studied. These visualisations make the visual language of the nineteenth-century paper reappear at scale, but while retaining a place for human interpretation.

[36] M. H. Beals, 'Close Readings of Big Data', pp. 620–23.
[37] James Mussell, *The Nineteenth-Century Press in the Digital Age*, p. 61.

Experiences in interpretation

While applying the visualisation tools developed in this chapter during the case studies, this project gained valuable experience on the operation and interpretation of these visualisations. However, these experiences both fed back into the tool's development, and are so meta in nature that they relate less to the case studies and more to the tool itself, and are therefore best discussed here. The two most prominent of these are the limitations imposed by computational time, and the choice of number of newspapers collated in one graph. The first relates directly to the possibility of using these visualisations in an explorative context, the latter prompts questions about the spatial patterns in nineteenth-century newspapers. Additionally, the general experience this project had with its visualisation experiments mirrored those of its topic modelling attempts described in the previous chapter. The comments made in the previous chapter on the importance of code quality, version control, and planning of inputs and outputs apply to this side of the project as well. However, some experiences are more specific to visualisations than others. These are the higher demand on computational resources, the question of keyword selection, and the requirement to develop additional interpretative resources.

The first of these relates to, the much higher demands of the visualisations on computational resources compared to topic models, which significantly coloured the way in which they could be used. The first major limitation imposed by these computational requirements relates to the number of articles that can be visualised at one time. As described above, the visualisation process requires a large

amount of RAM, and more importantly, the process of gathering the required data takes a significant amount of time. This carries implications for the way in which the visualisation can be utilised, which has to be conservative. This means the choice of data to visualise has to be in the function of single questions, and using the least possible data to form an interpretable graphic; in practice, this meant only three keywords for five-year slices for only a pair of case studies could be done in a realistic timeframe. While this might eventually be solved by code optimisation, which would be the next step were this an actual tool development cycle, part of the intent of this project was to discover what the limits of a single-researcher project in tool development are: when building our own tools, we have to accept that they will have some limitations in efficiency that simply cannot be overcome, which force the researcher to default to smaller datasets and thus to answering more focussed historical questions. Additionally, even more so than with the topic models, the speed limitations have methodological implications. Selection of the articles to be visualised is even more critical and conservative: only the keywords that generate article collections that will certainly be useful in addressing the questions in the case studies can be visualised. This unfortunately means that using these graphics in an exploratory form is currently impractical.

Yet this was not the only interpretative issue that these visualisations encountered. One of the other key questions that needed addressing was the choice of source. Would it be more beneficial to visualise articles from one title only, or articles from across multiple newspapers. Both were found to have their advantages. When using just a single title, it was far easier to work out what a given space represented. For example, it was quickly discovered that the two right-most

columns on the last page of *Reynold's Newspaper* were the weekly economy and stock columns. Thus, the meaning of the concentration of imperial keywords appearing in that space was much more easily established. Additionally, these kinds of patterns in the use of its page-space have been theorised to represent an important part of the identity of a publication.[38] Visualising the articles from a single publication makes this pattern visible.

However, the use of articles from multiple newspapers means that we can learn about cross-industry practices in the placement of articles. For example, over five local and national weeklies, the Parliamentary news page(s) were all found to be either page 3 or 4. One key issue with this approach is the scale at which such a visualisation would need to be undertaken to be relevant. In order to make a founded claim about cross-industry practices it would need to visualise a large number of articles from an equally large number of papers. Such a visualisation would quickly run into the computational constraints of this version of the tool. Thus, both ways of visualising placement are capable of producing new insights into the source material, and which one is more appropriate depends on the research question that the visualisation is meant to answer. For the majority of the questions in the case studies, this thesis found it was more valuable to visualise individual newspapers and then compare the results manually. The ease of understanding the patterns that appear, and the certainty of knowing that these are the results of a single visual identity outweigh the possible insights from an industry-wide approach.

[38] James Mussell, *The Nineteenth-Century Press in the Digital Age*, p. 72.

Secondly, there was the question of page visualisation to consider; would the visualisations differentiate between pages, or would all pages be superimposed onto a single plane? Again, both approaches offer their advantages and disadvantages. When visualising article placement without regard for pagination, the result allows a study of the general themes of a paper's layout. This was found to be less relevant for Victorian papers, which follow a strict column-based layout, but it could be valuable for more modern newspaper studies. Visualising pages separately, when coupled with the single-title approach, means that specific repeating elements, weekly columns and opinion pieces could be identified and interpreted.

Based on the experiences of analysing the graphs generated by this tool, this thesis can recommend that the tools it has developed be used with a clear goal in mind, as they are not suited to exploratory research contexts due to their computational time requirements. It simply takes too long to produce a figure that may or may not be of use to the research question under investigation. Additionally, it advises the visualisation of articles from a single source and respectful of the article's pagination, in order to uncover repeating elements in the paper under investigation.

Conclusion

In conclusion, this chapter has set out the method this thesis developed for visualising article placement. Unlike previous methods it only relies on the non-textual spatial data that is encoded with the articles, and does not use the size of the articles as an analogue for importance. Unlike previous methods it also boasts a high level of scalability and a low reliance on human data tagging, while still retaining human authority in the analysis step. This method relies on the creative re-uses of pieces of the archive, which were never intended to be used as such. The visualisations show the density of articles that match a set of selection criteria on each page of the newspaper, which can then be linked back to certain categories of article through analysing these specific spaces. This chapter also gained insight into the way the development and design of this visualisation tool shapes its usefulness, for example by choosing a certain colouration for clearer analysis.

What is particularly interesting in the case of visualisations of article placement, is that it fundamentally opens up new possibilities for research, some of which will be explored in the next chapter. Exploring article placement has only been done on a small scale, as it was time-intensive when using the prior existing methods. The development of this tool allows for overviews of article placement to be generated without human involvement. It makes it possible to answer questions about the space occupied by specific genres of content; it makes us ask whether these places were static over time or if articles changed places; it generates questions about the redesigning of the layout by incoming editors. It has the potential to be valuable for both the field of periodicals studies and for historians

wishing to gauge the impact or prominence of certain content. However, in its current embryonic form, it suffers from a lack of secondary literature and historical theory to embed itself in. At present, there has been very little work done on the visual language of Victorian newspapers, the way their layout spoke to their readers. Did readers value the front page the same as we do nowadays, or did the presence of the advertising wrapper mean the key content would be most eye-catching on the inner folio? Did readers learn to expect certain patterns to their newspapers that editors themselves were restrained by, such as expecting the Parliamentary debates to be on page three? There are countless questions like this, but most of the answers are yet to be found. Additionally, theories on content placement in a historical context will need to be developed; those that are available for newspapers generally consider only more modern columnless newspapers, which makes them unusable for content such as that in this archive.

Chapter 5: Imperial Identity as Case Studies

In the final part of this thesis, the methods developed in the previous chapters will be applied to historical study. This will show their effectiveness, their concrete value as historical tools, and their limitations. The topic of these case studies will be the expression of imperial identity in the British press during the nineteenth century. More specifically, they will explore how the press aided in constructing and propagating an imperial identity to its readers, the majority of whom were unlikely to come into direct contact with the empire themselves. Imperial identity makes for an ideal case study for topic modelling and visualisation tools. Firstly, it is at the same time very nebulous and obvious in its presence: while on one hand there are a range of clear markers for the presence of empire, mainly in the shape of geographic names, there are also many ways to invoke it without the signal being so clear-cut. For example, the keyword 'empire' on its own has a variety of uses, only a few of which signify the presence of the British empire or are meant to invoke feelings of imperialistic pride overtly. More common are covert uses of the term, such as the names of ships or theatres, which are not directly related to the imperial project. Meanwhile, in diametric use it can signal the presence of an imperial contender, such as 'the French Empire'. Each of these, in its own way, is useful for learning about the ways people interacted with the empire. This means that the advantages of topic models and article placement visualisation over keyword searches can be clearly demonstrated. The second reason for imperial

representation as the case study of choice is that there already exists a substantial body of work on the subject, which provides the theoretical framework within which to explain the findings.

This framework rests on the concept of 'banal imperialism', an offshoot of identity- and nationalism studies. Banal Nationalism, which will be explored in detail later, considers national identity to be constructed and maintained by subconscious, daily reminders of the nation-state. These are called 'flaggings'. For example, the Union Jack flown atop the Cunard Building on the Liverpool waterside flutters without any ceremonial presence – it simply reminds passers-by of their national identity. When using this conceptual framework to study the British Empire in newspapers, it makes sense to search for expressions of imperial presence that meet two criteria: (1) they have to be common and repeated – weekly adverts for a household good are considered flaggings, while a one-off column on the Sepoy rebellion is not; and (2) they have to be covert in their imperial message, in the sense that they are not aiming to evoke emotions – 'Rule Britannia' would not be considered a flagging, while a throwaway joke in a play might be.

Theories of Imperialism, Nationalism and Identity

Before we embark on our case studies, there has to be some discussion on relevant theory to contextualise their content. This has to be historical (or at least historically useful) theory, as we are now going beyond the process of devising the methods by which we could analyse and employ topic models and on to applying them on historical material with a historical research question. It is thus appropriate to discuss the theories of nationalism and imperialism that contextualise this

project's case studies. It will first cover nationalism and national identity, as the theory this project bases itself on was originally developed for those fields. This is particularly relevant, as it allows us to understand the role of newspapers in the creation and transmission of these identities. Subsequently, it will outline those theories and debates that have shaped the historiographical context of these case studies, on the nature and presence of imperialism in the British Isles. Discussing these bodies of literature enables the findings of the case studies to be grounded in historical debate.

Theories of Identity and Nationalism

The discussion of identity as a historical concept is mostly framed in the context of particular groups, and focuses on communal, rather than individual, expressions of identity. The leap from personal identities to group identities is highly debated, as it requires the crossing of a disciplinary boundary. Whereas personally identity is traditionally the field of psychology, group identity is more of a sociological matter. In addition to these groups, both personal and group identity have been increasingly explored by –mainly postmodern– philosophers, who tend to focus on language and discourse. This means there are many methodological debates that have to be resolved, in addition to bridging the theoretical gap between the self and the group.

However, this thesis is at present unconcerned with theories that cover *both* the self and the group, as it will focus solely on the latter. While there are several theories on group identity that could be applied to this topic, there is one that, while it has been further refined over the years, still lays the groundwork for most

of them. This is Anderson's imagined communities theory, which he first articulated in his 1983 book *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. The theory quickly found traction amongst historians as an explanation for the social changes in the nineteenth and twentieth centuries. Written at a time when the first cracks were starting to show in the Marxist theory of historical development, his ideas were expanded upon by several other authors, and his book was revised and reprinted twice, in 1991 and in 2004. There are two key concepts which Anderson defines that will be used in this thesis: the link between official nationalism and imperialism, and the imagined community as a vessel for group identity. The link between official nationalism and imperialism is laid explicitly by Anderson. He theorises the birth of colonial nationalism as the result of 'pilgrimages', (meta) physical journeys undertaken by a separate middle class, from the periphery to the centre. However, they will not be able to go to the centre directly, as a result of the barriers that the colonising state has laid down, and thus they will, on their journey, meet others like themselves. This leads to a feeling of connectedness not just between the people that physically meet, but also with all the others that try and reach the centre.[1]

Anderson theorised that because of this, it was inevitable that colonies developed their own imagined national community, as nationalism and imperialism are, at least from the perspective of the colonized, fundamentally incompatible. The example Anderson uses to illustrate this is the policy of 'Macaulayism'. While the goal of this educational policy was to create a native administrative class,

[1] Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, Rev. and extended ed. (London: Verso, 1991), pp. 54–60.

"Indian in Blood and colour, but English in opinion, morals and intellect", later referred to as 'the Palls', it led to a group disconnected from the masses they were supposed to govern, but also restricted from moving either vertically towards the centre or horizontally to other colonies.[2] This led to the imagining of a community amongst these "strangers in their native lands", which in turn gave rise to nationalism in the colonies.[3] The key concept throughout the book is the importance of this imagined community as a vessel for identity, which Anderson equates with nationalism or a national Identity. He does not explicitly develop the applicability of the Imagined Communities theory beyond the national, although he does state several examples of non-national imagined communities, if only to investigate why these did not translate into national feelings. The real beauty of his theory, however, is that it defined its key component –the imagined community– broadly enough so that it is applicable outside of the study of nationalism alone. "It is *imagined* because the members of even the smallest nation will never know most of their fellow-members, … yet in the minds of each lives the image of their communion. … it is imagined as a *community*, because, … the nation is always conceived as a deep, horizontal comradeship."[4]

This theory was enthusiastically adopted, especially by historians. Today, it enjoys such popularity that it has "become commonplace to observe [nations] are in part imagined communities."[5] This is mainly due to its lack of jargon and its

[2] Lord Thomas Babbington Macaulay, 'Minute on Indian Education (1835)', in *Archives of Empire: From The East India Company to the Suez Canal*, by Mia Carter and Barbara Harlow (Durham & London: Duke University Press, 2003), I, 127–38 (p. 137); Anderson, pp. 91–94.
[3] Anderson, pp. 90–94.
[4] Anderson, pp. 6–7.
[5] David Cannadine, *Ornamentalism: How the British Saw Their Empire* (London: Penguin Books, 2001), p. 3.

conceptual simplicity, as well as its broad applicability and versatility. It also works well with other theories popular amongst historians, such as Hobsbawm's Invented Traditions.[6] The latter allows for the inception of shared rituals that create connections between its participants, the former allows for those ritual and traditions to gain meaning in the propagation of the national idea. This ties into complexity theory: a small and local ritual may have large and wide-ranging effects, while a large gesture may cause only minor changes in a population's feelings.[7]

The theoretical framework that this project relies on most was formed on the intersection of 'imagined communities' and 'invented traditions': Michael Billig's theory of Banal Nationalism. Banal Nationalism was coined in 1995 to explain the influence of the less overt acts of nationalism on the creation and maintenance of the national imagined community. Take, for example, the flag. Billig identified a strict difference between the flag which is meant to *signal* something and is consciously noticed, like the flag on Remembrance Day; and the *symbolic* flag which is unconsciously seen and forgotten about, but omnipresent, like the tricolore on loaves of French bread. He is especially interested in the latter, as they are "providing banal reminders of nationhood: they are 'flagging' it unflaggingly. … The remembering is mindless, occurring as other activities are being consciously engaged in."[8] Of particular interest to this project is the role of newspapers in this 'flagging' of the national identity. In this respect, Billig places specific importance on the concept of deixis, the way words are used context-

[6] Eric Hobsbawm, *The Invention of Tradition* (Cambridge: Cambridge University Press, 1983).
[7] Eric Kaufmann, 'Complexity and Nationalism', *Nations and Nationalism*, 23.1 (2017), 6–25 (p. 8) <https://doi.org/10.1111/nana.12270>.
[8] Michael Billig, *Banal Nationalism* (London: SAGE, 1995), p. 41.

specifically. Examples include "*we* have decided to leave the EU" or, "*this* great country", referring in this context to 'all Britons' and 'Great Britain'.

Billig observes that this is especially prevalent in newspapers, which flag the nation on a daily basis. He notes that another way this 'flagging' takes place is by the placement of articles within the pages of the papers. The reader is expected to know the 'domestic' from the 'foreign' news, which reinforces the distinction between the own group and the outsider.[9] With this in mind, the project will try to determine if there is an *imperial* identity that is flagged in these Victorian newspapers as well as a national one. With this notion, Billig refers back to earlier work on identity formation building on social identity theory, which noted the importance of the (perceived) difference between the ingroup, or the group the individual belongs to, and the outgroup, those that do not belong to 'our' group. It noted that people tried to maintain a positive group identity by making favourable comparisons between the ingroup and the outgroup.[10] This was later applied in an imperial context by Edward Saïd as the 'othering' of non-Europeans.[11]

In conclusion, the theoretical foundation of these case studies rests on two bases: Banal nationalism, and, by extension, imagined communities. Most of the case studies will benefit more from exploration through the lens of banal nationalism, as it offers more practical handles to understand the language that may be found in newspapers. It is beyond doubt that the empire will be flagged; outside

[9] Billig, *Banal Nationalism*, pp. 116–24.
[10] Bethan Benwell and Elisabeth Stokoe, 'Theorising Discourse and Identity', in *Discourse and Identity* (Edinburgh: Edinburgh University Press, 2006), p. 25.
[11] Edward W. Said, *Orientalism* (London: Penguin, 1991).

of the press, there are more than sufficient examples of imperial flags in street names and music-hall songs. The goal of the case studies is not so much to explore one imagined community to the fullest extent, but instead to search for the various ways in which such communities may have connected with the imperial project or had an imperial dimension. Over the course of the case studies, we will look at the intersections between the empire and sports fans, women, and members of the military, amongst others. It is likely that some of these imaginary communities overlap, as for example, a Lieutenant in the army may also be an avid follower of horse-racing. Thus, we are looking for the way in which all these smaller communities collectively understand the imperial community, as it is flagged through the newspapers. But to do so, this thesis also has to understand the historiography of empire, as it provides the context in which we can interpret the sources.

Historiography of the British Empire

The historiography of the British Empire is extremely complex; ever since the nineteenth century, historians have tried to understand it. Providing a comprehensive overview would be impossible. Instead, this section will provide an overview of some of the debates in the history of imperialism on the British Isles. These are needed to properly value the achievements of the case studies; while they are not intended to be the main contribution to knowledge of this thesis, they will highlight new evidence for some of these theories by virtue of testing new methodologies.

In organising this wide body of historiography, the schema observed by Paul Ward is very useful, though it needs some supplementing. He observes three schools of imperial historiography: The Manchester School of cultural history, The Marxist-Socialist School, and the superiority school.[12] This last one, led by P.J. Marshall, which perceives imperialism as an outlet for pre-existing tendencies or a sense of uniqueness and superiority already existent in British society before the 1870s, is less useful for this particular thesis, as it is impossible to prove or disprove based on the source material this project uses; in order to be effective in such an argument, it would need to include a longer period before 1870 than just 20 years. Therefore, this thesis substitutes this strand of historiography with a school of traditional ('Little England') historians who wrote mainly during the first half of the twentieth century and for whom the empire was self-evident. This group has since evolved to take a position in the debate on the morality of empire vis-à-vis the Saidist and New Imperial History groups, and contains authors who argue that, on balance, imperialism was to the benefit of those colonised. Many of these authors see no need to investigate the deeper reasons for imperialism, because as they see it, there is no responsibility for its actions that needs to be accounted for.

The 'Manchester School' of imperial studies is mainly defined by the work of John McKenzie. Scholars of this persuasion look for the ways in which imperialism dominated British cultural life, particularly after the 1870s. The debate on the meaning of empire to the average citizen in Britain came to the fore in 1984 with the publishing of J. MacKenzie's *Propaganda and Empire: The Manipulation of*

[12] Paul Ward, *Britishness since 1870* (London: Routledge, 2004), pp. 15–16.

British Public Opinion 1880-1960. In this work he challenged the prevalent idea that the British were indifferent to imperialism, because of the diffuse nature of the British imperial experience. Instead, he set out to show the prevalence of empire in everyday life in Britain, whilst still acknowledging that "imperialism meant different things to different people in different times".[13] He observed that this view was the result of Imperial History focusing on Britain's impact on the empire, while the smaller body of work that did look at imperial influences on Britain considered imperialism a matter of the elites.[14] MacKenzie showed that traces of imperial ideology could be found throughout popular culture of the time.[15]

With his work he paved the way for like-minded academics, such as John Springhall and Penny Summerfield, to investigate imperialism in the broadest sense of the word. The 1986 collaboration on imperialism and popular culture provided many researchers with a starting point to look for imperialism in places that had previously been overlooked.[16] Semantically, these approaches required a broader definition of empire and imperialism than had been in use at the time. To make their research possible the definition of imperialism was expanded to cover such topics as art, music-hall entertainment, juvenile fiction, and marketing.[17] These historians view Victorian and Edwardian Britain as a society steeped in imperialism

---

[13] John M. MacKenzie, *Propaganda and Empire: The Manipulation of the British Public Opinion 1880-1960* (Manchester: Manchester University Press, 1984), p. 1.
[14] MacKenzie, *Propaganda and Empire*, pp. 1–3.
[15] MacKenzie, *Propaganda and Empire*, pp. 253–58.
[16] *Imperialism and Popular Culture*, ed. by John M. MacKenzie (Manchester University Press, 1986).
[17] John O. Springhall, '"Up Guards and At Them!": British Imperialism and Popular Art, 1880-1914', in *Imperialism and Popular Culture*, by John M. MacKenzie (Manchester: Manchester University Press, 1986), pp. 49–72; Penny Summerfield, 'Patriotism and Empire: Music-Hall Entertainment, 1870-1914', in *Imperialism and Popular Culture*, by John M. MacKenzie (Manchester: Manchester University Press, 1986), pp. 17–48; J. S. Bratton, 'Of England, Home, an Duty: The Image of England in Victorian and Edwardian Juvinile Fiction', in *Imperialism and Popular Culture*, by John M. MacKenzie (Manchester: Manchester University Press, 1986), pp. 73–93; MacKenzie, *Propaganda and Empire*, pp. 130–36.

(whether it was consciously aware of it or not) and the British Empire as a major component in its self-image. It has found popularity amongst followers of Edward Said, who were looking for signs of 'orientalism' in the past.[18] Deeply influenced by the 'cultural turn' of the 1970s and 1980s, this renewed focus on culture opened the field of imperial history to researchers in cultural studies, and saw many historians change their approach to historiography.[19] These scholars produced analyses of British imperialism in all things cultural, most prominently novels.[20] Many of them hailing from a different discipline, their methods and practices sometimes led to harsh criticism from more traditional historians, including from MacKenzie himself.[21]

One of their main points of critique is the former's great reliance on the works of the post-modern philosophers Foucault and Derrida, which has its own jargon that is not immediately apparent to those not intimately involved with it. In addition, it occasionally relies on debatable readings of historical texts, without providing supporting evidence from sources. Simultaneously, a new generation of historians was influenced by post-colonialism and post-modernism but took into account criticism of the cultural theorists, and as a result they formed the school of 'new imperial history'. This school tends to focus on the voices that history forgot and, to overcome the difficulty in researching these groups through

[18] Edward W. Said, *Culture and Imperialism* (London: Vintage, 1994).
[19] Peter Burke, *What Is Cultural History?*, 2nd edn (New Jersey: Wiley, 2008).
[20] Graham Dawson, *Soldier Heroes: British Adventure, Empire, and the Imagining of Masculinities* (London: Routledge, 1994).
[21] Bernard Porter, *The Absentminded Imperialists: Empire, Society and Culture* (Oxford: Oxford University Press, 2006), pp. xx–xxi; John M. MacKenzie, *Orientalism: History, Theory and the Arts* (Manchester: Manchester University Press, 1995), p. 214.

established methods, took to more culturally-focussed approaches.[22] The cultural history approach to imperialism that MacKenzie introduced is currently one of the more dominant in the field. For a somewhat more modern example, Dane Kennedy's *Britain and Empire* covers not only the economic and political history of British imperialism, but also considers its role in popular entertainment, the press and sports.[23] It also finds its use in researching the imperial dimensions of almost every part of everyday life, from food to money and from sports to professional associations. These case studies are sometimes presented in the form of books, but their primary means of dissemination is in journal articles.[24]

Second, historians like Richard Price and David Porter approach imperialism from a more Marxist-Socialist point of view by focussing on the class relations that underpin imperialist feelings. Imperialism thus becomes a middle-class venue for social cohesion vis-à-vis organised labour movements and the aristocracy. This theme was developed by Bernard Porter in *The Absent-Minded Imperialists* (2004), a book which reinvigorated the debate even before it was released.[25] He argues that just as domestic and foreign policies differ, so does the face of the government. Whereas to the rest of the world, Britain was defined by and identified herself with her empire, to the average Britton it was defined as a moderate, democratic nation, which just so happened to have an empire. Also, as

---

[22] *A New Imperial History: Culture, Identity and Modernity in Britain and the Empire 1660-1840*, ed. by Kathleen Wilson (Cambridge: Cambridge University Press, 2004).

[23] Dane Kennedy, *Britain and Empire, 1880-1945* (London: Longman, 2002). 60-62; 93

[24] Susie Protschky, 'The Colonial Table: Food, Culture and Dutch Identity in Colonial Indonesia', *Australian Journal of Politics & History*, 54.3 (2008), 346–57; W. Mwangi, 'The Lion, the Native and the Coffee Plant: Political Imagery and the Ambiguous Art of Currency Design in Colonial Kenya', *Geopolitics*, 7.1 (2002), 31–62; Simon Glassock, 'Good Sports? Scotland, Empire and Rugby c.1924–1928', *Sport in History*, 36.3 (2016), 350–69; Matthew C. Hendley, *Organized Patriotism and the Crucible of War: Popular Imperialism in Britain, 1914-1932* (Montreal: McGill-Queen's University Press, 2012).

[25] Porter, *Absentminded Imperialists*, pp. xx–xxiii.

the empire was maintained by only a handful of people, and participation in this maintenance was strictly regulated by social class, it required of the general population nothing but apathy. Porter proposed to look at the various classes instead of the nation in general, and pointed out that the degree of imperial commitment, as well as their conception of empire, differed greatly.[26] Therefore their degree of identification with the empire also varied greatly. He makes the distinction between imperialism (the dominating and aggressive drive for conquest) and imperial culture (what has also been called indirect imperialism), and by doing so could argue that Britons were not imperialist, but were steeped in imperial culture.[27] One other notable historian of this school is David Cannadine. His book *Ornamentalism: How the British Saw their Empire*, approaches the empire as "having been the vehicle for the extension of British social structures, and the setting for the projection of British social perceptions".[28] These social structures in Britain are hierarchical classes, while within the empire there existed a parallel one: race. However, the empire should be viewed as "first and foremost a class act, where individual social ordering often took precedence over collective racial othering."[29]

Finally, there is the 'little England' school, which developed into a modern whiggish strand of imperial history. In its original form, the Little Englanders were fronted by A. J. P. Taylor. One of the most prolific British historians of the 20th century, his style and methods influenced generations of English-speaking historians. He is singled out by MacKenzie as responsible for creating "a school of

[26] Porter, *Absentminded Imperialists*.
[27] Porter, *Absentminded Imperialists*, pp. 319–20.
[28] Cannadine, *Ornamentalism: How the British Saw Their Empire*, p. xix.
[29] Cannadine, *Ornamentalism: How the British Saw Their Empire*, pp. 6–8.

'Little Englander' historians which saw imperialism as essentially an irrelevance to domestic British history."[30] While it has to be noted that Taylor never said anything of the sort explicitly, it is understandable how MacKenzie arrived at that statement. Taylor was a traditionalist historian whose seminal work on the time period, *The Struggle for Mastery in Europe* (1939), is considered a product of 'Pure Diplomatic History'.[31] It focussed on the actions of the great European powers during the 19th century and approached it almost solely from the perspective of international politics.[32] Whenever colonialism or imperialism was mentioned, it was always in the context of some European event or a struggle for power between European states, an approach that was copied by his followers. He also approaches the colonies and their relation to Britain in a similar vein in his later *English History 1914-1945* (1976).[33] Incidentally, the only time Taylor focussed on colonial history, he wrote that "the average Englishman was ashamed of the Empire", which implies there was at least a broad awareness of imperialism, even though he does not qualify any of those terms.[34] In general, the 'Little Englander' school saw the empire as irrelevant or as self-evident.

In this category we also find earlier historians who wrote substantial bodies on Imperial history, such as Ramsay Muir (1922) and Paul Knaplund (1942). To these historians, the empire was an everyday fact of life, and any reading of its

---

[30] MacKenzie, *Imperialism and Popular Culture*, p. 2.
[31] Donald J. MacLauchlan, review of *Review of The Struggle for the Mastery of Europe 1848—1918*, by A. J. P. Taylor, *Weltwirtschaftliches Archiv*, 79 (1957), 66–68; Edward B. Segel, 'A. J. P. Taylor and History', *The Review of Politics*, 26.4 (1964), 531–46 (pp. 540–41).
[32] C. J. Wrigley, *A.J.P. Taylor: radical historian of Europe* (London: I.B. Tauris, 2006).
[33] A.J.P Taylor, *English History 1914-1945*, The Oxford History of England, XV, 3rd edn (Oxford: Oxford University Press, 1976).
[34] A.J.P Taylor, *Germany's First Bid for Colonies, 1884-1885: A Move in Bismarck's European Policy* (Basingstoke: Macmillan, 1938), p. 24.

presence in British society, if mentioned at all, began and ended with Britain being an imperial nation. In Muir's *A Short History of the British Commonwealth*, he puts the focus on the history of England first and foremost: it is not until 1,326 pages into the text that he focusses on Britain's colonies, and then only for a single chapter. Earlier he describes the conquest of India as an "amazing and undesired dominion", stating that "India had been rescued from the anarchy of the eighteenth century by the rise of British power".[35] Knoplund, writing two decades later, concedes that "It was a famous dictum, … that the British Empire was founded in a fit of absence of mind. … Empires, especially of the size of the British Empire, represent the presence of mind of somebody."[36] However, the index to his work still reads like a who-is-who of imperial heroes such as Dalhousie, Kitchener and Gordon. Overall, these historians paid little attention to the social or cultural aspects of imperialism, seeing it solely as a political project.

While this approach to imperial historiography fell out of favour post-war as social and cultural history became of greater importance, the tone of this early imperial historiography carried over to a new generation. This mainly concerns itself with the morality and post-facto accounting of the empire. The former, named after Edward Said of the Orientalist school, fundamentally hold the premise that the British Empire was a vehicle for oppression and exploitation. The study of historical imperialism then becomes a question of collective accountability. All

[35] Ramsay Muir, *A Short History of the British Commonwealth*, 4th edn, 2 vols (London & Liverpool: George Philips & Son, 1927), vol. I, pp. 812; II, 544.
[36] Paul Knaplund, *The British Empire, 1815-1939* (London: Hamish Hamilton, 1942), p. v.

the aforementioned strands fall within this group, each presenting their own reasoning for what factor was responsible for the growth of imperialism in Britain.

Conversely, there exists a school of conservative, anti-revisionist historians who see the empire as, fundamentally, a project of civilisation and working towards a 'greater good'. They feel no need to explore deeper reasons for imperialism, as it carries no responsibility. Jan Morris, for example, closes his trilogy on the British Empire (1979) with "When I began to write the book I thought I was describing something definitive in human history, but I have ended it seeing the imperial story in gentler but nobler terms, as a flicker of the divine progress. … The arrogance of the Empire, its greed and its brutality was energy gone to waste: but the good in the adventure, the courage, the idealism, the diligence had contributed their quota of truth towards the universal fulfilment".[37] Lawrence James states in his conclusion: "On the whole, Britain's Empire was a moral force, and one for the good".[38] A modern author in this school is Niall Fergusson, who begins the conclusion to his study of imperialism with:

> Of course no-one would claim that the record of the British Empire was unblemished. … Yet the nineteenth-century empire undeniably pioneered free trade, free capital movements and, with the abolition of slavery, free labour. It invested immense sums in developing a global network of communication. It spread and enforced the rule of law over vast areas. Though it fought many small wars, the empire maintained a global peace unmatched before or since. In the twentieth century it more than justified its existence, for the alternatives of German and Japanese rule were clearly far worse.[39]

[37] Jan Morris, *Farewell to Trumpets: An Imperial Retreat*, Pax Britannica, 2nd edn, 3 vols (London: Faber & Faber, 2012), III, pp. 558–59.
[38] Lawrence James, *The Rise and Fall of the British Empire*, 2nd edn (Abbacus: London, 1998), p. 638.
[39] Niall Ferguson, *Empire: How Britain Made the Modern World* (London: Penguin, 2008).

Works like these show that the curtain of imperialism still has not fully drawn on British historiography, and that the case studies undertaken by this thesis are useful in potentially providing further proof of the existence of imperial sentiment in Britain. If these case studies can show a widespread network of imperial flaggings in the line of Billig, then it offers a vessel by which to critique the underlying assumption in many of these revisionist texts that the empire was a project of elites. Such widespread mentions of imperial locations and events would indicate that the imperial project was supported, at least tacitly, by a large portion of the population.

This project is inherently shaped by the work of John MacKenzie and the Manchester school, who provided the first looks at the prevalence of "imperial propaganda" in the aftermath of 'new imperialism' in the 1870's, and the impact this widespread propaganda had on the British middle class. Further work by Bernard Porter provided additional nuance by further exploring the element of class, mainly the feelings of solidarity between members of the working class within the empire. Finally, Paul Ward applied Michael Billig's theory on Banal Nationalism to the British Empire, linking the formation of a national identity with the pageantry and pomp of imperial celebrations.[40]

However, while all of these historians mentioned above employ newspapers as evidence, their use is peripheral. As argued by Michael Wollf, periodicals should be at the forefront of cultural history and the history of ideas.[41]

[40] Ward, *Britishness since 1870*, pp. 12–18; Billig, *Banal Nationalism*.
[41] Wolff, 'Charting the Golden Stream'

Some historians have engaged more closely with the relationship between the press and the empire, and due to the nature of the sources they work on often rely on thematic or title-specific case studies.[42] This is an approach that this thesis follows in its own choice of case studies. Readers were reminded on a daily basis about the empire and their place in it, and read on their daily commutes of the trials and triumphs of the British Empire – their empire – these expressions become the nexus through which imperialism propagates.[43] In a similar way as Facebook and Twitter serve as the lenses through which the world is observed and understood in the twenty-first century, so did the periodicals and broadsheets of the nineteenth century. One facilitated Brexit, the other the Crimean War. Only through placing periodical sources, with all their peculiarities and idiosyncrasies at the centre stage can the modern historian understand the mindset and worldview of the Victorians.

Case Study One: Establishing a Baseline

The key question then becomes: how can a researcher working 150 years after the fact possibly make sound discoveries about the layers of identity that existed in research subjects that are unavailable to interview? The initial admission must be that we cannot. However, we can study the language they used to describe events and the world around them. There has been extensive research into language and identity on living subjects, starting with the work of Benveniste, who developed the notion of subjectivity in language. The speaker, and his identity, could

[42] Chandrika Kaul, *Reporting the Raj: The British Press and India, C. 1880-1922* (Manchester University Press, 2003); Simon J. Potter, 'Empire, Cultures and Identities in Nineteenth- and Twentieth-Century Britain', *History Compass*, 5.1 (2007), 51–71 <https://doi.org/10.1111/j.1478-0542.2006.00377.x>; Simon J. Potter, 'Jingoism, Public Opinion, And The New Imperialism', *Media History*, 20.1 (2014), 34–50 <https://doi.org/10.1080/13688804.2013.869067>.

[43] Simon J. Potter, *News and the British World: The Emergence of an Imperial Press System, 1876-1922* (Clarendon, 2003), pp. 9–11.

transform language from an abstract system to a system of communication.[44] As a field, discourse and identity studies has established itself over the last quarter century, and closely allied itself with theories of social constructionism. These hold that the social world is not an objective reality, but a constant process of construction by human action. Thus, identity is not a concept that an individual has, but what they construct through social interaction.[45] In practice this social interaction takes the shape of language, both written and spoken. It is on this realisation that this thesis builds. Regardless of whether language reflects, conveys, or constructs identities, or does all those at the same time, language is shaped and changed by identity. All language is thus in some way the language of identity – but some language more than others. What makes topic models useful in the study of this language is that they can categorise banal mentions of a concept that is both a nexus for identity, such as 'France' or 'empire' or 'enemy', but which also occurs too often for a human researcher to close read in all its contexts. This ability is particularly useful when dealing with sources that use the chosen language in a wide range of linguistically distinct texts, such as news sources.

However, no study can rely only on news sources, as the agenda-setting powers of the press invite the risk that the historian takes the 'reality from above' for the 'reality from below'. The Victorians themselves were well aware of these issues, and several commentators at the time warned of the power of the proprietor over the content of a paper.[46] In order to avoid accidentally investigating the

[44] Emile Benveniste, 'Subjectivity in Language', *Problems in General Linguistics*, 1 (1971), 223–230.
[45] Anna De Fina and Sabina Perrino, '"Transnational Identities"', *Applied Linguistics*, 34.5 (2013), 509–15 <https://doi.org/10.1093/applin/amt024>.
[46] Michael Bromley and Tom O'Malley, 'Introduction', in *A Journalism Reader* (Psychology Press, 1997), pp. 13–15.

imperial identity of the editor or proprietor rather than the reader, the historian has to look at the news, content, and language that has slipped beneath the editor's hand: the everyday and the banal. Family notices, market news, shipping and railway timetables, even weather reports. These everyday articles can tell us much about the presence of imperial identities in newsreaders, as they offer a particular organisation of the world to the reader. For example, do the economic pages list imperial companies alongside British ones? These articles also feature a limited and distinct vocabulary that is highly repetitive. This means such texts are identified by topic models with a relatively high degree of accuracy. Thus, using topic models, we can gain an insight into the scale on which these banal articles circulated, and calculate their relative importance. This will be the goal of the case studies presented in this chapter: to explore the banal flagging of the empire by newspapers in a variety of subjects.

Before this project can make a judgement on the topic composition of the case studies, it needs to establish a baseline for comparing the other topic models against. In other words, this thesis needs to answer the historical question: what was the composition of an average newspaper between 1850 and 1900? Without an answer to this question, it is difficult to draw any conclusions from the topic models or the visual analysis of any of the subsets generated by keyword search. A complete analysis of every article in the dataset is beyond the computational resources of his project. Fortunately, Elasticsearch offers the option of generating

a consistently random sample based on a user-provided seed.[47] This means that while the sample is random, it is also reproducible, and thus remains available for future researchers. A random sample of 10,000 articles per year was modelled in the same way as the keyword-selected subset, by decade-long intervals with N_T = 20. These topics were then labelled with the general labels that were based on experience with the topic models: economy (ECON), trade (TRDE), politics (POL), crime (CRIM), adverts (ADV), Welsh news (WELSH), and personal news (PERS), as well as one for any topics that did not fit in this schema (MISC). The topic decomposition of the decade-long slices showed little variation, thus these were averaged to one for expediency. This is presented in Figure 5.1.

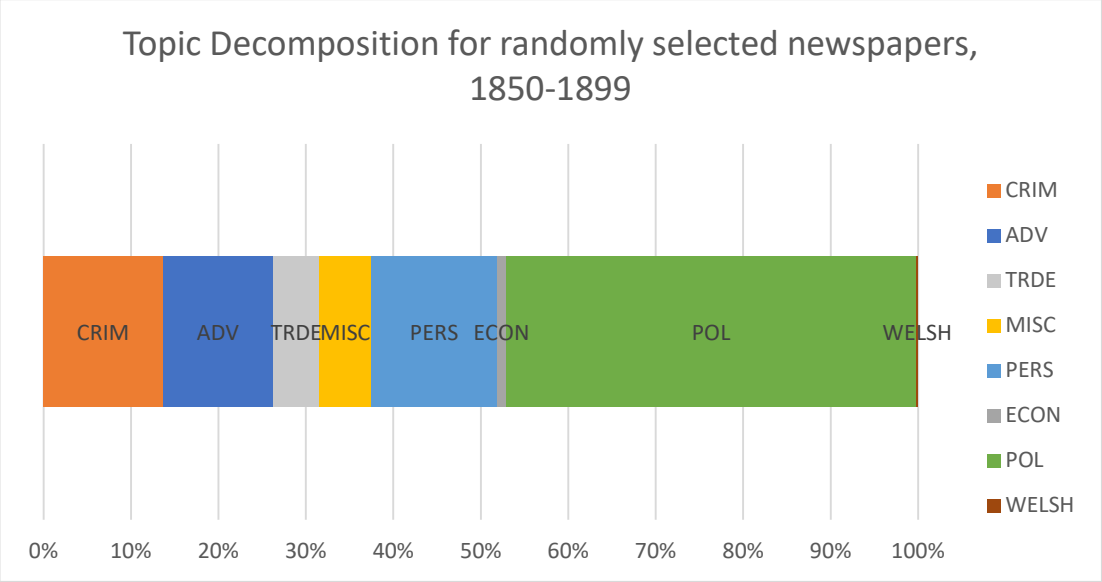


Figure 5.1 Topic Decomposition of Randomly-selected newspaper articles 1850-1900. The most dominant category of topic is the Political, which represents nearly half of the content. Criminal news, Advertisements, and Personal news each contribute between 12 and 15% of topics. Welsh articles represent less than 1% of the set.

[47] 'Random Scoring | Elasticsearch: The Definitive Guide [2.x] | Elastic' <https://www.elastic.co/guide/en/elasticsearch/guide/current/random-scoring.html> [accessed 1 February 2019].

The topic decomposition suggests the predominance of political reporting in the nineteenth century press, as close to half the articles appear to be political reports. The categories were verified by close reading twenty articles per topic per decade. Additionally, looking at the general content maps developed for understanding the visual placement analysis also pointed towards a dominance of political news. One major contributing factor to this high percentage of content appears to be the Parliamentary reports; because topic models rely on linguistic difference between texts, the words that only occur in political texts cause a strong correlation to be inferred. Terms such as 'honourable', 'lord' and 'bill' specifically identify Parliamentary reporting.

In a similar way, while the Welsh newspapers represent only a minor fraction of all news, these are the most coherent group of all: the difference in language between them and other articles is so pronounced that assignment to this topic comes with a large degree of certainty. The minute percentage of Welsh articles also conforms to the expectations of a random sample. There are only two Welsh-language newspapers in the archive for the period under investigation, *Baner Cyrmu* and *Ganaedl*, thus only a minor contribution of Welsh-language articles to the random sample is to be expected. The contribution of trade and economic news together comes to 7%, with the inclusion of texts such as railway timetables which are not economic per se but are indicative of economic activity. Personal news, adverts, and crime news each come in at between 12 and 15% of the corpus. Finally, the miscellaneous category, which is comprised mostly of (specific forms of) OCR errors fills out the remaining 6%.

One of the main insights gained from this topic model relating to this chapter is the lack of an imperial topic forming at any of the settings used. There are two possible explanations for this. It is possible that that imperial reporting was small scale and limited to incidents and crises, and that the volume of imperial news was simply too small for it to form into a topic. The problem with this theory is that there are simply too many mentions of imperial markers that appear with a simple keyword search for this to be true. This seems to suggest a second explanation: that at the level of detail offered by 20-topic models, the empire is not signified by language that is sufficiently distinct from the other topics. That is, an article on the economy of India has, from a linguistic point of view, more in common with an article on the economy of Yorkshire than it does with an article on the politics of another imperial territory like Canada.

This exposes one of the major downsides to topic modelling for historical research: it is impossible to tell beforehand at what number of topics a topic that addresses the research question will form, or whether they will form at all. To re-use the analogy of the professor and their research assistant, we don't know how many boxes we have to supply before one of them contains *only* receipts for last year's conference. To further complicate the situation, in our case, it is possible that the topic we need will never appear, as it may remain hidden in the clutter of OCR noise. If this is the case, it shows the need to perform keyword searches before topic modelling the outcome, as it is the only way to steer the models towards answering a specific question. It also means that we cannot rely on topic models alone to provide structure to the search for imperial identities and imperialism.

Fortunately, other historians have already developed frameworks that may be used. In his 2004 study of British national identity, Paul Ward identifies five 'facets' through which to view and understand British national identity prior to 1960: monarchy and empire, gender, rural versus urban tensions, popular culture and leisurely pursuits, and political debates and divisions.[48] Each of these can correspond to one or more of the labels that were observed above, or to these categories appearing in specific keyword-induced contexts. For example, the personal column of family notices could indicate a gendered dimension, or one of social class.[49] This thesis will attempt to use Ward's facets as the main framework to explore the imperial identities not only post- but also pre-1870. Where appropriate, it will draw on other theoretical frames.

In itself, this case study has given us a first insight into the power of topic models. It has established the topic decomposition of the archive, providing the first view at the archive's profile. It has verified the expected dominance of political news, and has been able to quantify the relative amounts of certain news genres. It has also established that imperial reporting is not linguistically distinct enough to form as a separate topic.

Case Study Two: Economic News

In its first case study, this project will explore the economic topics that appear in the topic models. The connection between Britain and its colonial possessions has been widely theorised as economic in nature, and the appearance of these topics

[48] Ward, *Britishness since 1870*, pp. 10–12.
[49] Kristy Hess and Sarah Pinto, 'Forever In Our Hearts', *Media History*, 26.2 (2020), 105–21 (pp. 109–11) <https://doi.org/10.1080/13688804.2018.1482205>.

allows for this connection to be explored.[50] It will look at the two topics that appear which relate to economic life: economic news and trade bulletins. Each of these collects a specific genre of news report. The economic news topic collects stock-, share-, and currency prices from across the world. The trade bulletins deal with much more tangible goods, often grains like wheat and barley, but also livestock and raw cotton. While both represent a very different economic experience, combined they are as crucial to our understanding of business in the nineteenth century as they were to businessmen of the time to operate in an imperial economic system.[51]

In order to make this an imperial exploration, we have to create a subset of articles by keyword search before applying the topic modelling tool. As the topic modelling takes care of the economic aspect of the question, all that is needed is to find a word that will nearly exclusively occur in colonial context. The keyword chosen for this subset was 'India', as outside of banal imperial occurrences (for example 'India Quay'), it exclusively refers to the British dominion. A topic model of up to 100,000 articles per decade at $N_T = 20$ was constructed using the topic modelling pipeline described previously in chapter 3. Special interest during the annotation phase was given to the various economic sections and topics that emerged. The resulting topic model is shown in Figure 5.2.

[50] See for example: Bernard Semmel, *The Rise of Free Trade Imperialism: Classical Political Economy the Empire of Free Trade and Imperialism 1750-1850* (Cambridge: Cambridge University Press, 1970); Nancy Fowler Koehn, *The Power of Commerce: Economy and Governance in the First British Empire* (Cornell University Press, 1994); D. K. Fieldhouse, 'Gentlemen, Capitalists, and the British Empire', *The Journal of Imperial and Commonwealth History*, 22.3 (1994), 531–41 <https://doi.org/10.1080/03086539408582938>; Philip Williamson, *National Crisis and National Government: British Politics, the Economy and Empire, 1926-1932* (Cambridge University Press, 2003).

[51] Heidi J. S. Tworek, 'Political and Economic News in the Age of Multinationals', *Business History Review*, 89.3 (2015), 447–74 <https://doi.org/10.1017/S0007680515000677>.

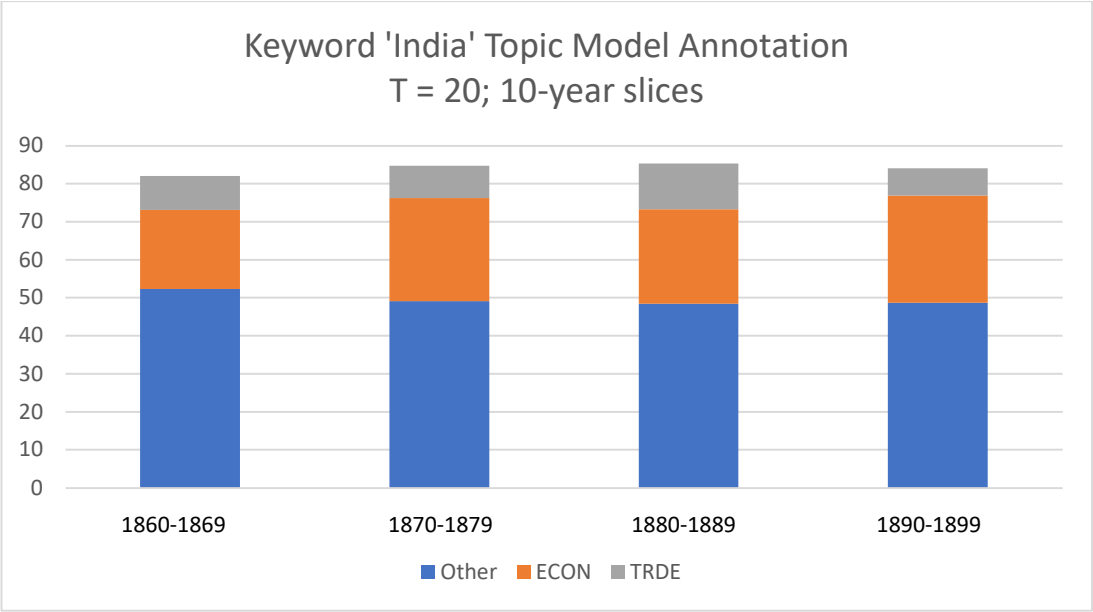


Figure 5.2 Economic and Trade News in Imperial reporting. The topic decompositions show Economic and Trade news represented a substantial amount of imperial reporting, and that the amount of such articles remains relatively constant throughout the century.

The topic decomposition clearly shows the dominance of economic news in the subset. We may interpret this as a significant portion of the discourse surrounding India in British newspapers being economic in nature. The Anglo-Indian economic link is flagged in approximately 30 percent of the corpus every decade, if we take the sum of econ and trade. As shareholding was a widespread middle- and upper class pursuit, the economic sections are of particular interest to these readers, particularly if they either did business in India or, more likely, held a few share- or stock options in companies that did.[52] This dominance of news on the stock market needs some further investigation, as this shows the different readership of a Victorian paper. The latter half of the nineteenth century saw a

[52] Mark Freeman, Robin Pearson, and James Taylor, "'A Doe in the City": Women Shareholders in Eighteenth- and Early Nineteenth-Century Britain', *Accounting, Business & Financial History*, 16.2 (2006), 265–91 <https://doi.org/10.1080/09585200600756282>; Graeme G. Acheson and John D. Turner, 'The Secondary Market for Bank Shares in Nineteenth-Century Britain', *Financial History Review*, 15.2 (2008), 123–51 <https://doi.org/10.1017/S0968565008000139>; Janette Rutterford and others, 'Who Comprised the Nation of Shareholders? Gender and Investment in Great Britain, c. 1870–1935', *The Economic History Review*, 64.1 (2011), 157–87 <https://doi.org/10.1111/j.1468-0289.2010.00539.x>.

major expansion of investments by all who could afford them, as economic expansion seemed limitless and therefore the risks involved were perceived as small. Especially after the 1855 Limited Liability Act and the Companies Acts of 1856 and 1866 provided both a legal framework for such companies, and additional protections for small investors against unscrupulous fraudsters, Britain became a "nation of shareholders".[53] By 1900, roughly two-fifths of the national wealth was tied up in company shares, and large numbers of the upper- and middle classes relied, at least in part, on the income that these paid out in dividends and interests.[54] This development was reliant on the spread of accurate financial information, which newspapers provided.[55]

However, in the close reading of the topics, we also find topics that hint at a close rural-imperial relationship. In the 1870-79 model, one topic of 0.18% of the subset contained articles from rural newspapers reporting on literal cross-pollination between Britain and India. It reported the potential of cross-breeding Indian and British tobacco plants, in order to create a species that was more resistant to draught and disease, and provided a higher yield.[56] As Britain did have a domestic tobacco production, such a breed would be of great interest to growers

[53] George Robb, *White-Collar Crime in Modern England: Financial Fraud and Business Morality, 1845-1929* (Cambridge University Press, 2002), pp. 26–28; 3; Donna Loftus, 'Capital and Community: Limited Liability and Attempts to Democratize the Market in Mid-Nineteenth-Century England', *Victorian Studies*, 45.1 (2002), 93–120; For a background of the British Financial system at the time, see: Mary Poovey, *The Financial System in Nineteenth-Century Britain* (Oxford University Press, 2003).

[54] David C. Itzkowitz, 'Fair Enterprise or Extravagant Speculation: Investment, Speculation, and Gambling in Victorian England', *Victorian Studies*, 45.1 (2002), 121–47 (p. 121).

[55] Mary Poovey, 'Writing about Finance in Victorian England: Disclosure and Secrecy in the Culture of Investment', *Victorian Studies*, 45.1 (2002), 17–41 (p. 17).

[56] 'Political and Social: Notes and Comments', *The Examiner* (London, 2 February 1878), pp. 10-11 (138-139).

in the UK.[57] The topic also collected articles on seed enhancement and cross-pollination of imperial and domestic crops. So, in a way the relationship between the colonies and the rural population was still very pronounced: the seeds they sowed benefitted from improvement by cross-breeding with colonial species.

In another article in the same topic, the *Sheffield & Rotherham Independent* reports a lecture on the way failures in British agricultural policy have facilitated the 1876 Indian Famine, and that Britain had a moral duty to take better care of its Indian subjects.[58] In his lecture the speaker invokes a comparison to the British agricultural laws, such as the Enclosure Acts and the Corn Laws. These had been in effect since the late eighteenth century and were intended to protect the interests of the land-owning classes by respectively making it easier to evict tenants and keeping the price of grains high. Campaigning against these laws had been long and politically heated, managing repeal of the latter as part of a broader package of reforms in 1846; the matter of evictions from enclosed land would continue until 1882.[59] The audience would have been familiar with the impact these laws had on the everyday life of people subject to them, and the intent of the speaker was obviously to generate a sympathetic reaction from the audience towards the Indian farmers. This shows the way that even in peripheral urban centres, that is, towns

[57] Jordan Goodman, *Tobacco in History: The Cultures of Dependence* (London and New York: Routledge, 2005), pp. 6; 139–43.
[58] 'Our Government of India, and the Famine', *Sheffield & Rotherham Independent* (Sheffield, 25 August 1877), p. 2.
[59] See for more details on these laws: William D. Rubinstein, 'The World Hegemon: The Long Nineteenth Century, 1832 - 1914', in *A World by Itself: A History of the British Isles*, ed. by Jonathan Clark (London: Pimlico, 2011), pp. 451–565; Christopher Harvie, 'Revolution and the Rule of Law (1789-1851)', in *The Oxford Illustrated History of Britain*, ed. by Kenneth O. Morgan, 2nd edn (Oxford: Oxford University Press, 2009), pp. 419–62; Jeffrey G Williamson, 'The Impact of the Corn Laws Just Prior to Repeal', *Explorations in Economic History*, 27.2 (1990), 123–56 <https://doi.org/10.1016/0014-4983(90)90007-L>; Michael Turner, 'Enclosures in Britain 1750–1830', in *The Industrial Revolution A Compendium*, ed. by L. A. Clarkson, Studies in Economic and Social History (London: Macmillan Education UK, 1990), pp. 211–95 <https://doi.org/10.1007/978-1-349-10936-4_4>.

outside of the major cities, the population engaged with the way agriculture was run outside of their personal sphere of interest.

The appearance of India in these contexts supports the interpretation that British interests in India were predominantly economic in nature. But it also links two elements that Colley argues also shaped British identity during the Georgian period – an identity that these trade bulletins twice affirm. First, the notion of Britain as an Imperial nation, unified at home against threats abroad, a project in which the Irish, Scottish and English could jointly work towards a goal greater than their old national allegiances.[60] Second, the deep-seated belief that the British Empire was distinct from others, as it was an 'Empire of Trade', was reaffirmed by these notices.[61]

This notion of Britain as a champion of free trade and the marketplace of the world had been present since the middle of the eighteenth century, and while the relationship with India had become increasingly administrative rather than commercial in that century's closing decade, it still persisted.[62] The recognition of the depth of, and imbalance in, Anglo-Indian economic integration began during the nineteenth century, led by the work of Indian nationalists and scholars such as Dadabhai Naoroji and Romesh Chandra Dutt. They concluded the economic relationship between the colony and the centre was inherently exploitative and was responsible for the wealth of India 'draining away' to Britain.[63] Rebuttals to this

[60] Linda Colley, *Britons: Forging the Nation 1707-1837*, 2nd edn (New Haven and London: Yale University Press, 2014), pp. 118–21; 145–47.
[61] Colley, *Britons*, pp. 99–101.
[62] David Cannadine, *Victorious Century: The United Kingdom, 1800-1906* (London: Penguin, 2018), pp. 49–51.
[63] Dadabhai Naoroji, *Poverty of India* (London: Vincent Brooks, Day and Son, 1878); Romesh C. Dutt, *The Economic History of India Under Early British Rule: From the Rise of the British Power in 1757, to the Accession of Queen*

thesis were produced by imperialist defenders, such as Edward Thornton, E.H. Nolan and R. Knight. The latter said in 1866, during the preamble for a debate on Anglo-Indian financial relations: "My judgement is clear that whether we have regarded (sic) to the comparative, or positive, defects or merits of our rule, it is entitled, upon the whole, to a favourable verdict".[64]

However, this interpretation of the exploitative economic relationship has been held as orthodoxy in historiography since the 'Orientalist turn' of the 1980s. Since then, there has been more research into this area, focussing on the legal and infrastructural consequences of British rule.[65] In its most extreme form, the profits from India, both by being the biggest importer of British produce and biggest supplier of raw materials (mainly cotton and coal), are argued to have propped up British financial world supremacy. "For practical purposes, the British had annexed the parts of the Indian economy that could strengthen their place in the global economy."[66] The results of this case study so far underwrite this conclusion. The British papers we have looked at so far highlight a strong linkage with India, in both the financial stock-and share categories, and the goods trade. Especially the latter, while less voluminous in the amount of articles dedicated to it, shows very close connections between the centre and the colony.

Victoria in 1837, II vols (s.l.: K. Paul, Trench, Trübner & Company, Limited, 1902), I; Romesh C. Dutt, *The Economic History of India in the Victorian Age: From the Accession of Queen Victoria in 1837 to the Commencement of the Twentieth Century*, II vols (s.l.: K. Paul, Trench, Trübner & Company, Limited, 1904), II.

[64] Robert Knight, *The Indian Empire, and Our Financial Relations Therewith: A Papaer Read Before the London Indian Society* (London: Trubner & co., 1866), p. 3.

[65] Tirthankar Roy, *Economic History of India, 1857-1947* (Oxford University Press, 2011) <https://ideas.repec.org/b/oxp/obooks/9780198074175.html> [accessed 22 January 2019].

[66] Darwin, pp. 187–88.

When looking at more local papers, this conclusion holds. Closely investigating a single local title, we find the *Preston Chronicle* contains a significant amount of economic news from the empire, as well as large numbers of adverts. However, using this particular paper in the archive has to be done carefully; the British Library seems to have treated the *Preston Guardian* as an alternative title for the *Preston Chronicle*, whereas they were in reality two different papers. However, a sampling of the scans in the archive for the period 1850 to 1893 shows the title of *Chronicle* is correct, despite the information given through the archive interface and supporting material. This is predominantly of interest as the two papers had very different readerships. The *Chronicle* was the paper for the liberal members of the middle class and advocated the Whig political cause; it was therefore in an adversarial relationship with the *Guardian*, which catered to a more Tory, working class audience.[67]

The first observation that can be made from the annotated topic models is the prevalence of the empire in advertising. For every decade, between 25 and 30% of the mentions of India, Canada or Australia come from adverts. The theme that is becoming apparent in this thesis is that these adverts are a prime source for banal interaction with the empire, as they provide the tiniest flagging, a short moment of remembrance as long as it takes the eye to scan a column, that the empire is British, and that the reader is part of it. The nature of these adverts varies wildly, from real estate in Canada ("more properties in the Dominions available upon request") to Indian Teas, and from ladies clothing ("Wedding Trousseaux, Skirts & Dressing

[67] Andrew Hobbs, *A Fleet Street in Every Town: The Provincial Press in England, 1855-1900* (Cambridge: Open Book Publishers, 2018), pp. 45–47.

gowns, Indian Outfits") to miracle pills supposedly "recommended by the personal surgeon to the viceroy of India".[68] Adverts show that on a daily basis, for the entire second half of the nineteenth century, the reading public of Preston was reminded of their position as consumers in an interconnected imperial system of trade, just as their work in the cotton mills made them part of that system as producers.

This trend is also true in other local papers, such as the *Hampshire Telegraph* (average 29.8%) and the *Huddersfield Chronicle* (average 42.3%). This shows the way that the empire was flagged in the daily mindset of the British reader as an economic entity. The economic link was also underlined by the reporting on the various market prices of goods throughout the empire. For example, the *Huddersfield Chronicle* dutifully reported on the price of wool in Australia and Canada, the *Hampshire Telegraph* published the price of shipping goods from one place to the other, and the *Preston Chronicle* covered the cotton price in India compared to Louisiana. The interesting realisation here is that all these are connected to the economy of the local area that the paper serves, showing that a British worker or entrepreneur was very aware of the fact that he was not only supported by the empire, he was also in competition with it.

In conclusion, the analysis of the amount of different kinds of economic news has led to a confirmation of the economical connections between Britain and India, as expected based on the historiography. The economic columns of both national and local newspapers were spaces in which the empire could be flagged

[68] 'Adverts and Notices', *Preston Chronicle* (Preston, 18 November 1882), p. 4; 'Walker & Sons. - India, China & Ceylon Teas', *Aberdeen Journal* (Aberdeen, 22 July 1898), p. 4; 'Ladies Outfitting, Oldham & Sons.', *Freeman's Journal* (Dublin, 30 March 1877), p. 1; 'Small Dose, Small Pill, Small Price: Clarke's Pills', *Western Mail* (Cardiff, 21 August 1890).

on a daily basis, represented as an integrated part of the globe-spanning imperial trade system. These topics highlight the nature of the imperial project as economic and widespread: everyone was part of the British Empire, farmers and bankers both stood to gain or lose from fluctuations on imperial markets. This initial case study has also shown that the use of a topic model for historical research can be highly beneficial to research questions that benefit from classification of texts into different categories. By clustering the archive without any human biases, they essentially provide a fresh pair of eyes. For human historians investigating the encounters with India the average Victorian reader had in his morning paper, stock bulletins would most likely be dismissed as 'noise' and receive at most a minor mention. Even if they were investigated further, making an accurate measurement of exactly how often India appeared in this particular context would have been significantly more difficult without the use of topic models.[69] By allowing a number to be put on their significance as a percentage of the subset, the argument that there existed a significant difference in the quantity of news relevant to urban readers versus that relevant to rural readers when it came to India can be strengthened.

However, this case study has also begun to highlight some of the downsides to topic models, mainly that they cannot be relied on to generate that categorisation that the researcher desires. When setting the number of topics, there is no way for the researcher to know which categories will result, nor is it evident how increasing or decreasing the number of topics will cause different categories to form. Thus

[69] See for one example of this: Poovey, 'Writing about Finance in Victorian England'.

there is an element of chance involved in the research process: if a category does not appear, it may be due to a real lack of material for that topic to form around, or it may be because the settings were not quite right – and there is no way to know which is true.

Case Study Three: Family Notices and Gender

The second case study this project will undertake is investigating the personal news category, in order to discover if this category within the topic models can help us gain new insights in the closeness with which people experienced the empire. As this topic comprises mainly content that was generated by the readers themselves, with a desire for it to be read by others, it allows for an insight into the personal impact the empire had. This case study will do so by investigating the family notices topic that formed in the subset generated by the keywords 'India', 'Canada' and 'Australia' with $N_T = 20$. A breakdown of the percentage of this topic for each decade is presented in figure 5.3.

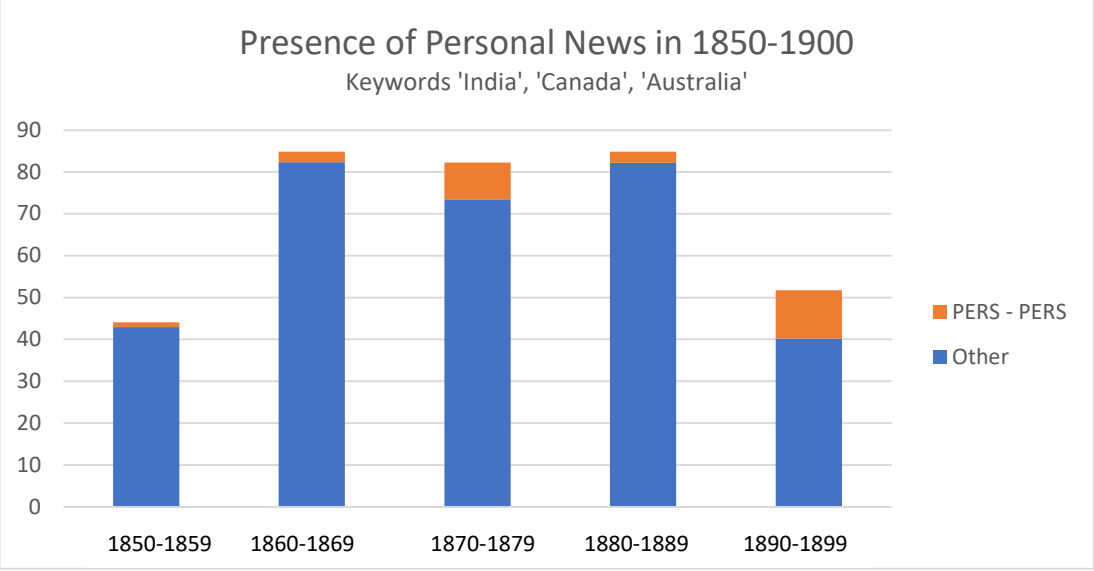


Figure 5.3 Percentage of personal adverts and family notices in subset generated by imperial keywords. The percentage of this topic is relatively small with two major peaks in the 1870s and 1890s.

These family notices were discovered to be a highly persistent topic, which appeared for nearly every keyword modelled. These notices were presented as lists, and often only one or two items in the article signalled back to colonial involvement. For example, in the *Preston Chronicle* of November 30th 1850, as part of a long list of marriages, it mentions that: "at Berhampore, Mrs. Anne Rea, aged 34, [married and became] wife of Wm. Rea, Esq., merchant, [of] Aska, East India, and [she is the] eldest sister of Mr. C. G. Hill, artist, of this town;".[70] These nuptials, taking place half a world away, were made relevant because the brother of the bride was known in the community. The only unknown here is if the notice was posted by the happy couple themselves or if Mr. Hill was informed of his sister's wedding by letter or telegram and had the notice placed himself. In any case, these notices represent the presence of the empire in extremely personal events.

Before undertaking this case study, I hypothesised that it would reveal a steady increase of family notices from the colonies being printed in British newspapers as the result of emigres wanting to keep relatives in the motherland informed of significant events in their family lives. However, the number of family notices does not match the trends of migration. While between 1850 and 1900, 1.1 million people migrated from Britain to India and Australia, there are major fluctuations on a decade-by-decade basis that are not represented in the numbers of notices.[71] These fluctuations may be explained because throughout the second

[70] 'Family Notices', *Preston Chronicle* (Preston, 30 November 1850), p. 3.
[71] *The Nineteenth Century*, ed. by Andrew Porter, The Oxford History of the British Empire, 5 vols (Oxford: Oxford University Press, 1999), III, fig. 2.5.

half of the nineteenth century, migration patterns to the colonies changed significantly.

While previously those moving to the colonies in general, and to India from Britain in particular had been almost exclusively male, improvements in medicine and technology lessened the perceived risks involved, and a larger portion of women travelled overseas. There was such a need for a support network that would enable women to undertake the journey that between 1849 and 1900 three different societies were founded to provide such schemes; the British Ladies Female Emigrant Society, the Female Middle-Class Emigration Society, and the Woman's Emigration Society, whose programmes were often sponsored by colonial governments.[72] Exact numbers are difficult to determine, for a single year over 300 single women arrived in Adelaide alone.[73] The creation of international networks of women led to the entanglement of the family and the empire.

The empire became a 'Family Affair', and with the transposition of family units from Britain to the colonies came the desire to keep the home front informed about intimate goings-on.[74] Such communication also helped to "produce personal forms of colonial knowledge for those who remained in the metropole", which

[72] Carol M. Martel, 'British Ladies Female Emigrant Society', *Historical Dictionary of the British Empire* (Westport, CT: Greenwood Press, 1996), 189–90.
[73] Brooke Weber, '"A Mad Proceeding": Mid-Nineteenth-Century Female Emigration to Australia' (unpublished PhD Dissertation, Royal Holloway, University of London, 2018), pp. 143; 288–91 <https://pure.royalholloway.ac.uk/portal/files/31011848/2018weberbphd.pdf.pdf>.
[74] Lisa Chilton, *Agents of Empire: British Female Migration to Canada and Australia, 1860-1930* (Toronto: University of Toronto Press, 2007), p. 21 <http://ebookcentral.proquest.com/lib/edgehill/detail.action?docID=4672409> [accessed 22 March 2019]; Elisabeth Buettner, *Empire Families: Britons and Late Imperial India* (Oxford: Oxford University Press, 2004), p. 4.

helped to further imperial awareness of those that stayed.[75] However, while this is likely to be a factor, it would still result in a steady increase in the number of family notices posted from the colonies, which is not what the topic model suggests.

We can also try to explain the amount of family notices by looking at the history of communications technology. The easier it becomes to communicate home, the more people will be able to place a notice, up to the point that communication becomes so easy they can simply contact just those they wish to notify. As the century progressed communications between the metropole and the peripheries were constantly improving. The years between the laying of the first transatlantic cable in 1865 and the first trans-pacific cable in 1902 were marked by a flurry of imperial telegraphic connections being made. Cables to India (in 1865), the Cape (in 1879) and Australia (in 1872) ensured that all colonial governments were in constant contact with – and under constant supervision of – the Colonial Office in Whitehall.[76] Britain especially wished for an imperial network which did not need to pass through other, potentially belligerent, countries.[77]

The connectivity had grown to such an extent that on the morning of her diamond jubilee in 1897, before she set out on her procession, Queen Victoria

---

[75] Esme Cleall, Laura Ishiguro, and Emily J. Manktelow, 'Imperial Relations: Histories of Family in the British Empire', *Journal of Colonialism and Colonial History*, 14.1 (2013).

[76] 'The Indo-European Telegraph.; Interesting Details of Laying the Line Difficulties with the Natives. the Manufacture of the Cable. Conveying the Cable. the First Station. Among the Arabs. Coming to Terms" with the Sheiks. Natural Difficulties. a Surprise. Dangers of Mud. the Last Obstacle.', *The New York Times*, 26 March 1865, section Archives <https://www.nytimes.com/1865/03/26/archives/the-indoeuropean-telegraph-interesting-details-of-laying-the-line.html> [accessed 19 November 2019]; 'History of the Atlantic Cable & Submarine Telegraphy - Cable Timeline' <http://atlantic-cable.com//Cables/CableTimeLine/index1850.htm> [accessed 19 November 2019]; Sir Timothy Augustine Coghlan, *A Statistical Account of the Seven Colonies of Australasia, 1895-6*, New South Wales Bureau of Statistics and Economics (Sydney, Australia: Potter, 1896), pp. 192–93; David Cannadine, *Victorious Century: The United Kingdom, 1800-1906* (London: Penguin Books, 2018), pp. 480–81.

[77] Standage, pp. 102–3.

could use the telegraph room at Buckingham Palace to send her jubilee message through the Central Telegraph Office and on its way to all the corners of her empire. Within two minutes, it had passed Teheran on its way east, and by the time her carriage was on the Mall, it had been received in Ottawa, West Africa, the Cape, Malta, Cyprus and the Caribbean.[78] Access to these cables was, of course, not limited to officials and royals. For a price, everyone could send messages across the world – and the price of a telegram dropped as the century wore on.

Still, telegraphs remained expensive and for use in extremely important occasions for the classes of people that would provide the largest amount of bodies moving to the colonies. Thus, while sending a single telegram to inform family back home of a birth or death would be just about affordable, sending an individual message to every family member and acquaintance would be prohibitively expensive.[79] People would telegraph either the newspaper directly, or contact a family member who would place the notice in their stead, thereby reaching their entire social circle for the price of just a single message sent. Some evidence for this practice may be found in the byline to many of these sections, that a notification of birth, death or marriage must be verified by a local representative of the paper or a government official; evidence that they are aware they are serving as a 'signal booster' for such family news. While such an increase-decrease model comes closer to matching the observations in the topic model, it does not fit convincingly.

[78] Morris, II, pp. 21; 27.
[79] Standage, pp. 63; 114.

So what is going on then? Is there anything inherent to the data that may explain why it is difficult to come to conclusions about the amount of family notices in British newspapers? In regard to the archive, an inherent bias seems unlikely. All papers had sections dedicated to family notices, and while initially information would arrive first in port cities such as Liverpool, Newcastle and Portsmouth, giving the papers there a higher potential for containing such notices, papers from these cities do not suddenly appear in the archive between 1850 and 1900, keeping the composition of the archive constant. Additionally, any increase driven by technological innovations leading to a broader adoption of family notices from abroad would be explained as a finding, not an anomaly.

A second potential explanation is in the technical handling of the newspaper data. It was noticed while performing the analysis needed for this case study that occasionally, these personal ads collate into a single topic with other, more general, adverts and notices. This failure of the topic model to differentiate between articles and family notices is compounded by the failure of the article segmentation to do so. This article segmentation problem is a known issue in digitised newspapers, as it attempts to define where each article begins and ends. The ideal result of this process is illustrated in figure 5.4.[80] However, such an outcome on real data is rare, and often segments overstretch.

[80] Thomas Palfray and others, 'Logical Segmentation for Article Extraction in Digitized Old Newspapers', in *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12 (Paris, France: Association for Computing Machinery, 2012), pp. 129–132 <https://doi.org/10.1145/2361354.2361383>.

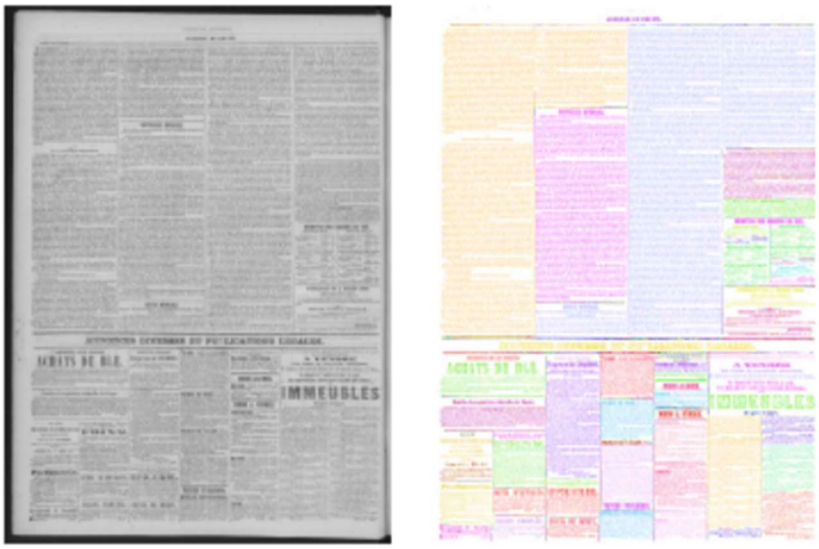


Figure 5.4 Example of automatic article segmentation. This example uses significantly more advanced techniques than were employed on the British Library Nineteenth-Century Newspaper Archive, but serves to illustrate the goal of the process. Palfray, pp. 132.

On several occasions, it was observed that if some other texts and family notices are taken as single documents, the family notices as a topic will be collated into that topic. This level of sensitivity of the model is illustrated strongly by the family notices and naval appointments blending together in the modelling of the *Hampshire Post and Telegraph*, where as little as sixteen failures of the article segmentation cause the family notices topic to be folded into naval appointments. We must also note that this might be intended by the editor: on an occasion where there were not enough family notices to make a separate section, they may have put them within the general adverts deliberately. This makes the topic modelling of these texts very difficult.

But despite these modelling difficulties, we can still learn from these family notices. Because their nature as personal notices intersects the domestic sphere, we may look towards them as a space in which we can see traces of women involved

in the empire. While little work has been done on the theoretical underpinnings of doing so, some authors have already used family notices to address questions of gender in colonial contexts.[81] The role of women imperialists specifically was crucial in the establishment of an effective British rule over its overseas possessions, as they – and the family units that came with them – provided the people who actually ran these far-off places.[82] The study of these family notices may then shine a light on the way the empire allowed the (re)negotiation of gender roles, which Ward presents as one of the facets of British national identity, in the context of an imperial identity.

Case Study Four: Military

The next case study that this project will undertake involves the way British newspapers reported on the British military. The armed forces have been theorised as important carriers of imperialistic sentiment.[83] The decision to investigate these aspects of imperial power is supported by Colley's work on British identity until 1837, she put forward the thesis that in order to forge an amalgamate of 'Britons' out of the Scots and English, there needed to be an external force to react against. "War and Empire were the means by which the union between Scotland and the rest of Great Britain was made real."[84] As military news can be nebulous and does not fall under one or two well-defined keywords, especially when it is investigated

[81] Hess and Pinto, p. 106.
[82] Lawrence James, *Raj: The Making of British India*, 2nd edn (London: Little, Brown and Company, 1998), pp. 221–23; *Britishness Abroad: Transnational Movements and Imperial Cultures*, ed. by Kate Darian-Smith, Patricia Grimshaw, and Stewart Macintyre (Melbourne: Melbourne University Press, 2007); Lisa Chilton, *Agents of Empire: British Female Migration to Canada and Australia, 1860-1930* (University of Toronto Press, 2007).
[83] See for case studies on this topic: John M. MacKenzie, *Popular Imperialism and the Military: 1850-1950* (Manchester University Press, 1992).
[84] Colley, *Britons*, p. 134.

in context of another such topic like the empire, this case study serves as an example of how topic models may help in historical research. The articles this case study collects could all have been collected through keyword searches, but that would have required an extensive list of all words that indicate solely the presence of the British military in an imperial context. For example, 'Burma' would only be a relevant keyword for those years that Britain was at war with the Burmese. This problem is even worse for the Royal Navy, who moved ships around the world continuously, and whose military commitment in imperial matters was often implicit and rarely involved actual firing guns in anger – just delivering the threat of doing so.[85] Without the option to keyword search, we have to rely on the ability of the topic modelling tool to create this kind of structure from general keywords.

In contrast with the case studies discussed above, this will not rely on keywords to create the imperial association before topic modelling takes place; instead it will topic model keywords that we expect to have some imperial association based on literature, and identify which topics within them flag the empire. This case study expects the discussions on imperial defence to use sufficiently different language from debates on national defence that they will appear as distinct topics.

For this, a selection was made of 379,939 newspaper articles that mentioned the word 'army', 194,853 that mentioned 'battle', and 493,257 that contained the keyword 'navy'; a small number of simple keywords with a large number of articles

[85] Lincoln Paine, *The Sea and Civilization: A Maritime History of the World* (London: Atlantic Books, 2013), pp. 547–49.

returned. These articles were modelled with $N_T = 20$, with a view towards testing Colley's thesis of War, and by extension Imperial War, being the flux in the forging of the national identity. If the military and the empire were closely linked, then we would expect one or several distinct imperial topics to appear within these articles.

The topic models of 'battle' unfortunately did not offer any insights in the role of the armed forces in (imperial) identity formation and propagation. The topic models that formed out of this dataset all covered either conflicts between foreign powers, or were non-relevant uses of the word, for example to denote a political or social conflict, or referring to historical battles – or the town near Hastings. In particular the American Civil War and the Franco-Prussian War were peaks for this keyword occurring. Instead, this case study will focus on the 'Army' and 'Navy' keywords, which correspond with the two branches of the armed forces at the time.

The Army

The topic model for the keyword 'army' can generally be split into two different kinds of topics. First there are the campaign reports, which are incidental reporting of the actions of British army units on operation. These cluster around wars, theatres and campaigns. For example, out of the twenty generated topics for the period 1850-59, twelve contain articles that relate to such conflicts, predominantly the Crimean War or one of the various colonial rebellions that western powers had to deal with. In the case of involvement of other nations' armies, these do not separate out, but cluster with the British forces they fight alongside. For example, one of the identifying terms for the Crimean conflict is 'Piedmont', a major partner

in the war. Though they arrived late, they distinguished themselves holding the besieging army's rear at the battle of Traktir Ridge in August 1855.[86]

Focussing in on the Crimean War a bit further, British readers could be expected to be well aware of the way the conflict was progressing; as noted by Markovits, the press not only presented a view from the front to the people at home, it also reported against the backdrop of the thousands of personal letters that would be sent home by the soldiers engaged in the long and static siege – which would subsequently be reprinted in British newspapers.[87] The topic models showed little in the way of these letters; it is likely that these would have been more numerous if other keywords had been used.

In later years, some very specific topics form around individual campaigns. For example, in the 1880-1889 model, a minor topic of 48 articles forms around the word 'Hazara'. These detail the brief campaign of the British forces to pacify the local population. In the decade earlier, the words 'Khyber', 'Afghan' and 'Russia' guide the topic of the Second Anglo-Afghan war, while the 1890-1900 model produces a specific topic for the Matabele War and one for the Siege of Chitral. These topics tend to cluster around belligerents and places, which is expected: these are the linguistic markers that signal articles about the same events.

The second kind of topic is the army list, which contains the promotions and awards given to individual soldiers and officers, as well as the articles listing of

[86] R.E. Dupuy and T.N. Dupuy, *The Collins Encyclopedia of Military History: From 3500 BC to the Present*, 4th edn (New York: HarperCollins Publishers Ltd, 1993), p. 906.
[87] Stefanie Markovits, *The Crimean War in the British Imagination* (Cambridge: Cambridge University Press, 2013), p. 42.

British army deployments and stations. Together these account for 18.4% of the corpus. These army listings form a regular feature of a paper, in which the various movements of the regiments of the British army (overseas) are reported for the benefit of those that are outside army life. They tend to be formatted as a list, and therefore are a clearly distinct type of content for the topic models to identify. The main beneficiaries of these articles were the families of those that served abroad, as they would be provided with a general update as to where their male relative was being sent. The meaning of this information for the reader underwent a distinct change during the nineteenth century, as Britain moved to its present regimental system. Before the 1868-1874 Cardwell reforms, a recruit could be placed with any army unit that had a need for men, without any personal choice in the matter, creating situations in which a nominal 'Highland' regiment contained more replacements from London, Manchester and Ireland than actual Scots.[88] Amongst other matters, the reforms kept the recruits from the same area together during their time in service, which was intended to create a close bond of camaraderie between the soldiers.

The result of this change means that a contemporary reader would have experienced the army lists in very different ways. Before the reform, it would have been a personal experience, where they would look for the one regiment in which their one family member served. Any kind of pride at the achievements of an army unit overseas would have been personal. After the reform, however, there is the potential for a togetherness in this reading to develop, as the reader would know

[88] Albert V. Tucker, 'Army and Society in England 1870-1900: A Reassessment of the Cardwell Reforms', *Journal of British Studies*, 2.2 (1963), 110–41 (p. 117).

that they are not just reading about the movement of their own son, but about the sons and brothers of their neighbours. This allows for a much more regional identity to express itself.

It should also be noted here that the army had a low social status throughout the nineteenth century, which only began to change in the middle of the 1880s. Considered 'unskilled workers' on relatively low pay, not allowed to marry, third-class citizens, and with little in the way of entertainment in garrison towns available for them, only the most desperate joined up; about a third of the yearly recruits had to be rejected for being undernourished.[89] However despised at home during peacetime, when they were successful in combat, they could fuel a local pride at 'our boys'.

These reforms were not easy for Cardwell to achieve, and his proposals sparked a flurry of debate in Parliament. This is reflected in the two topics that formed in the 60-69 and 70-79 slices, which are both characterised by the strong association of words like 'bill', 'committee', 'motion' and 'flogging'. This latter term may be strongly associated with the Cardwell reforms: amongst other matters, he sought to make army life more attractive, and thus gain more recruits, by clamping down on the worst excesses of military discipline. Flogging, in particular, which had been a common military punishment, was abolished as part of Cardwell's plans. An interesting note here is that the first years of reform are much more comprehensively discussed in the press, with 12.8% of the material for that decade in the topic, while only 5.6% makes up the 1870s slice. Thus, based on contextual

[89] Tucker, p. 136.

knowledge and topic models, we can narrow down the material that a researcher looking into these reforms has to sift through from 200,000 to just over 18,000. This is still an incredibly large amount, but a subsequent topic modelling step could reduce this even further.

After the Cardwell reforms, the army took on a much more explicit imperial dimension. As a consequence of his 'double battalion' reform, which amalgamated multiple regiments with only a single battalion into one regiment, the British army adopted the 'transit system' for foreign service. The army had been struggling with the demands for trained soldiers abroad since the dissolution of the East India Company Army in 1860 as the scale of colonial wars grew. The transit system created a conveyer belt of forces, with one battalion trained at home, deployed to Africa, then to India, to the Middle east, and back home; all the while the 'home' battalion acted as a reserve with which to replenish losses in the 'away' unit. The purpose was to create an explicitly imperial reserve, fit to wage a colonial war, and ready to defend the British Empire from any other European power's interference.

It needs to be noted though, that even after these reforms the British Army was sorely lacking compared to its contemporary continental counterparts. Field marshal Wolseley remarked that he did not know of a single battalion outside the Guards fit to go into the field against any European nation, but that was no concern, as it was not necessary for imperial defence.[90] The main issue was that even with Cardwell making a soldier's life more tempting, there was still a continuous shortage of soldiers, and both the Liberal and Conservative

[90] Tucker, pp. 38–39.

governments resisted implementing conscription. The upper ranks remained appointed by royal prerogative and filled with the offspring of the aristocracy and upper class. The true revaluing of the private soldier only came about during the Boer War in 1899-1900.[91]

We may thus conclude that there is evidence for a very personal connection to the empire in the press through the military troop movement lists. These allowed the relatives of soldiers and officers to keep abreast of their movements within the imperial system of troop rotations. While a career in the army was a last resort for many due to its low status and bad pay, this also meant that for the common soldier contact with home was rare, and these lists may be rare moments for his family to be informed of his whereabouts. For those without personal connections to the army, these lists may have invoked feelings of patriotic pride, similar to the feelings evoked by coverage of the Crimean war.[92] Further research would be needed, however, to verify this.

While unpredictable, the topic modelling also occasionally exposes the very explicit uses of the army in an imperial context. Its campaigns in India, Africa, and Afghanistan may form topics of their own, and while difficult to compare across models as they do not form reliably, they offer an insight in the sharp edge of the imperial project, relaying the bloody means by which its bounds spread "wider, still, and wider".[93]

---

[91] David Gates, *Warfare in the Nineteenth Century*, European History in Perspective, 7 (Basingstoke: Palgrave Macmillan, 2001), p. 180.
[92] Markovits, pp. 59–62.
[93] *Land of Hope and Glory* (London: Boosey & co., 1902).

The Navy

The topic models also find a collection of similar lists of navy postings. These are also published in a wide variety of papers, although they are far from the only mention. Topic decomposition suggests that the navy was much more in the public eye than the army ever was, by virtue of generating a much larger variety of topics. This observation corresponds with the assertion that Britain was primarily a naval power, both before and during the examined period of 1850-1900.[94]

One particularly interesting find is the use of the navy, most likely unofficial, in a variety of advertisements and product names. Navy Cut Tobacco is perhaps the most obvious in underlining its naval connection as a mark of excellence, but it was far from the only one. The May 1st edition of the *Hampshire Telegraph* in 1880, undoubtedly capitalising on the end of the academic term, contained numerous adverts such as this one for the diocesan grammar school in Southsea, stating that "In the above School the Sons of Gentlemen are received and thoroughly taught. More than Two hundred Pupils educated in this School have most successfully passed the Entry Examinations for the Royal Navy, Law, and Medicine…"[95] On the same page, an advert for Schweizers Cocoa brands itself as "widely adopted as mailed comfort in the Navy", while in the July 10th 1880 edition of the *Manchester Mercury* a Huddersfield instrument store drums up customers by stating it is "patronised by the Army, Navy, and Rifle Corps".[96] These advertisements show

[94] Rebecca Berens Matzke, *Deterrence Through Strength: British Naval Power and Foreign Policy Under Pax Britannica* (Lincoln and London: University of Nebraska Press, 2011); Roger Morriss, *Naval Power and British Culture, 1760–1850: Public Trust and Government Ideology* (London and New York: Routledge, 2017).
[95] 'Educational Advertisements', *Hampshire Post and Telegraph* (Portsmouth, 1 May 1880), p. 2.
[96] 'Advertisements and Notices', *Manchester Mercury* (Manchester, 10 July 1880), p. 7.

that the Royal Navy was a national institution that was often banally referenced for the readers, underlining their British identity.

The imperial connection is less evident in this data, which is unexpected. Many studies have been written exploring the connection between Navy, nation and empire; in particular how these various aspects developed throughout the nineteenth century.[97] Amongst others Jan Rüger, Lincoln Paine, John Wells and Ellie Miles all worked on or around this very broad topic using a variety of methodological approaches. Amongst these authors, there is consensus that the British Empire and its associated imperial identity has a strong naval component.[98] In a political sense, this expressed itself as a century-long 'two-navy standard' in which the Royal Navy was supposed to be as large as the next two largest fleets combined.[99] In a social sense, naval themes became a staple for fashion and theatre.[100]

However, it is important to note that the Royal Navy itself underwent a major change in the half-century covered by this case study, the significance of which cannot be overstated. We need to understand this change in order to understand the way references to the navy would be understood by readers of the time. To illustrate: in 1850, the navy began to plan the inclusion of auxiliary steam

[97] Jan Rüger, 'Nation, Empire and Navy: Identity Politics in the United Kingdom 1887-1914', *Past & Present*, 3.185 (2004), 159–87.
[98] John Wells, *The Royal Navy: An Illustrated Social History, 1870-1982*, New edition edition (Stroud: Sutton Publishing Ltd, 1996), p. 23.
[99] Paine, pp. 560–61.
[100] Amy Miller, 'Clothes Make the Man: Naval Uniform and Masculinity in the Early Nineteenth Century', *Journal for Maritime Research*, 17.2 (2015), 147–54 <https://doi.org/10.1080/21533369.2015.1094984>; Ellie Miles, 'Characterising the Nation: How T.P. Cooke Embodied the Naval Hero in Nineteenth-Century Nautical Melodrama', *Journal for Maritime Research*, 19.2 (2017), 107–20 <https://doi.org/10.1080/21533369.2017.1405632>.

engines for the wooden-hulled ships of the line then under construction, starting with *HMS Sans Pareil*; by 1900, the ideas that would lead to the design and launch of *HMS Dreadnaught* in 1906 had just begun to percolate in the naval community. *Dreadnaught* was the battleship that would make all others obsolete on her launch, driven by oil-fired steam turbines, armed with five 12-inch rifled guns and protected by up to 11 inches of steel armour.[101] The technological gulf is at least as significant as the one that separates Gutenberg's printing press from a laser printer. As an organisation, the navy underwent a similarly significant change, going from a highly conservative organisation to one that sought to be at the cutting edge of technological development.

This means that 'The Navy' possessed two distinct meanings to readers between 1850 and 1870, and 1870 and 1900. For the first period, it makes sense to draw on the work of Isaac Land, who studied the Royal Navy as a specific response to Linda Colley. He proves that the sailors of the Royal Navy were only slowly recognised as British, and that they mediated their identities for personal profit: choosing in which instance a national or imperial identity would offer them most benefit, not out of (mutual) loyalty to the nation.[102] 'Jack Tar' was a rough, uncouth and uncomplicated man. Consequently, invoking the navy would not have created a strong associative reaction amongst middle- and upper-class readers. Amongst

[101] Robert K. Massie, *Dreadnought: Britain, Germany and the Comming of the Great War* (London: Vintage, 2007), pp. 386; 468–79.
[102] Isaac Land, *War, Nationalism and the British Sailor, 1750-1850* (Basingstoke: Palgrave Macmillan, 2009), pp. 8–10; 166.

working-class readers, however, he was a figure that provided an opportunity for identification, as he was from a similar social strata.[103]

This contrasts strongly with the relationship between the navy and nationhood in the latter end of the period. From the 1870s onward, perception of the navy changes. Technological innovations and a desire by the Admiralty to be on the cutting edge of naval developments, meant that the new, modern sailor became a representative of late-Victorian and Edwardian values of masculinity. He was cast "as symbols of respectable British manhood celebrating their duty to nation and empire and their devotion to the family. The construction of the naval man's image as both a patriotic defender and dutiful husband and father stood in sharp contrast to the image of the brave but bawdy tar [before]".[104] The Royal Navy became a nexus around which imperial sentiments accumulated and coalesced, as its raison d'etre was to safeguard the lines of communication between Britain and her overseas possessions. Imperial unity became associated with a powerful fleet.[105] Newspapers in particular played a significant role in this process, by covering naval reviews extensively, and framing them as occasions on which the 'whole imperial family' could come together.[106] For example, *Punch* illustrated the 1897 Naval Review by showing the British lion rowing four cubs (representing

[103] Quintin Colville, 'Enacted and Re-Enacted in Life and Letters: The Identity of the Jack Tar, 1930 to Date', *Journal for Maritime Research*, 18.1 (2016), 37–53 <https://doi.org/10.1080/21533369.2016.1172840>.
[104] Mary Conley, *From Jack Tar to Union Jack: Representing Naval Manhood in the British Empire, 1870-1918* (Manchester: Manchester University Press, 2009), p. 3.
[105] Jan Rüger, *Great Naval Game: Britain and Germany in the Age of Empires* (Cambridge: Cambridge University Press, 2007), pp. 175–83.
[106] Rüger, 'Nation, Empire and Navy: Identity Politics in the United Kingdom 1887-1914', pp. 166–67.

Australia, Canada, New Zealand and the Cape) towards the assembled fleet in the distance.[107]

Looking back at the adverts that use the navy in their sales pitch, there is merit to this reading. While the percentage of the corpus made up of adverts remains stable around 20%, in the first two decades these are mostly for (life) insurance, or for public (bankruptcy) auctions. This latter is not surprising either: as the Royal Navy was on peacetime readiness, many of its officers were retained on reduced pay based on seniority, with the intent that they formed a reserve from which the navy could draw in the event of war. However, for many, this sum was not enough to live on, especially for those of junior ranks or those with short careers. The practice itself was discontinued in the 1860s, as rapid technological advancement meant that an officer that had been out of the navy for a decade would be just as far out of his depth on the new vessels as a raw recruit.[108]

By contrast, during the last three decades advertisements more along the lines of the ones discussed above, for services and luxury goods dominate. Yet more than commercial, the connection between the navy and the reader was personal, as the vast majority of mentions, approximately 40%, comes from lists of stations and promotions being publicised. These are almost exclusively reprints, as such news was originally published in the *London Gazette*, and was generally found in newspapers from towns with a strong military connection, such as Portsmouth or Huddersfield. While these only make up a small portion of each

[107] Rüger, 'Nation, Empire and Navy: Identity Politics in the United Kingdom 1887-1914', p. 177.
[108] N.A.M. Rodger, 'Commissioned Officers' Careers in the Royal Navy, 1690–1815', *Journal for Maritime Research*, 3.1 (2001), 85–129 (pp. 90–91) <https://doi.org/10.1080/21533369.2001.9668314>.

paper, they underline the military connection with the empire, as these notices mention whether an officer is posted to a base overseas. Moreover, they are regular features, which can be relied on to appear on a weekly or monthly basis.

What can then be drawn from the topic compositions generated from these newspaper articles about the armed forces? Firstly, that the majority of mentions come from the army and navy listings that the papers published for the family and friends of those that served. These offered an everyday reminder of their personal connection to the nation and the empire. For those who did not have this personal link, such articles still fulfilled a banal flagging of their national and imperial identity, by reminding the reader of the work done by those serving, wherever on the globe they may be – and whoever they may be, as just their shared nationality or belonging to the empire warranted their mention in these listings.

Secondly, they reflect the changing faces of the armed forces, both in the army and the navy. Through the efforts of reformers such as Cardwell and simple technological advancement, the two branches of the military became much more accepted into society. A clear indicator of this is the change in the volume of the army and navy listings, especially when they mention units sent abroad. For example, the *Preston Chronicle*, a liberal local paper, dedicated just 0.5% of its imperial news output to army lists in the decade 1850-59. By 1890-99, this had grown to 1.7%, a significant increase, which becomes even larger in absolute

numbers. In the 1880s the same paper begins to publish navy lists in addition to the existing army lists, buying into the 'cult of the Navy' observed by Tucker.[109]

The topic models have allowed this case study to rely on relatively general keywords when searching the archive. It has also shown that topic models can be used for addressing these kinds of questions, where there are no specific terms to guide a keyword search. It has also found the importance of contextual knowledge in understanding the results of a topic model historically. For example, knowing the way the public image of the Royal Navy changes between 1850 and 1900 informs very different analyses of newspaper adverts that appear as topics.

Case Study Five: Politics

This case study will investigate the political reports that are present in the corpus. Ward identifies the importance of political identities in the formation of Britishness, though these are predominantly national in nature: allegiance to the tribe of the working-class Labour party, middle-class Liberal party and the upper-class Conservative party. While the Labour party was not founded until 1900, the period under investigation is still home to political initiatives for the working class, therefore Ward's facet of 'politics' has merit.[110] In this case study, topic models are used to explore how politics were represented in the British press from an imperial context. It will do this by looking at the way the international political 'other' is reported on and compare this with the reporting of political news from the empire, through topic modelling and placement analysis. Essentially, the question is thus:

[109] Tucker, p. 111.
[110] See for a discussion of political engagement by the working class: Edward P. Thompson, *The Making of the English Working Class* (London: Penguin, 1963).

as a feeling of Britishness may follow from identification with a political group vis-a-vis others nationally, can there also be a feeling of imperial belonging following from an identification with a political group or union internationally? This question will be investigated using two different tools: the topic models that have been used in the preceding case studies, and the visualisation tools developed in chapter four.

Topic models and visualisation rely on a subset of articles, so we generate two for comparison, based on multiple keywords. India, Canada and Australia are used as markers of empire. These were chosen because they were three of the major possessions of the British Empire, spread around the three major continents on which it held territory, and they related to colonies in different stages of development and of different types: Canada as a developed settlement colony, Australia as a pioneer settlement colony, and India as the 'crown jewel' of the empire. Earlier tests to find suitable search terms for Britain's African holdings proved unsuccessful, as there the geographic descriptors were too fluid over the fifty-year time period investigated. While this issue was solved in the previous case study by topic models, the tool this thesis has developed does not allow topics to feed into further topic models.

For the comparison with foreign news, three imperial competitors were chosen. France, Britain's oldest enemy and competitor in Africa, Russia, which competed with Britain for influence in Central Asia, and Germany, which began to challenge British naval power towards the end of this period, but with whom Britain had strong dynastic ties in the earlier part of the century. These subsets were made for both the entire dataset and further filtered down for different

papers, allowing a closer look at individual newspaper titles. This latter subsetting is particularly relevant for the heatmapping of article placement. After deliberation, America was considered but not chosen as a potential competitor in the imperial sphere. While it was a source of economic competitive anxiety among the British public, its empire-building practices did not compete with Britain's.[111] The decision to join the articles collected by different queries was taken because the number of articles found by keyword search was too small to reliably model. On their own, some papers and periods had only 1,000 to 3,000 articles, while at the peak end, 10,000 to 14,000 articles were found by keyword search. Thus, by using several papers and keywords together, a subset size of at least 10,000 articles could be assured.

Topic Modelling

For its first tool, this case study will rely on topic modelling to investigate the context in which foreign and imperial markers were used. This is not easy, as there are significant issues with the dataset underlying this model. We have already explored the difficulty of subset size, and how this is addressed by collating subsets based on multiple keywords together. However, this still caused the issue that some of the smaller collections have a number of broken topics, where the same label can be applied to several topics, as well as cases where two or more topics are very close together and have very low article association scores. Still, it is better to model

[111] For background on American Imperialism, see: Walter L. Williams, 'United States Indian Policy and the Debate over Philippine Annexation: Implications for the Origins of American Imperialism', *The Journal of American History*, 66.4 (1980), 810–31 <https://doi.org/10.2307/1887638>; David Eric Brody, 'Building Empire: Architecture and American Imperialism in the Philippines', *Journal of Asian American Studies*, 4.2 (2001), 123–45 <https://doi.org/10.1353/jaas.2001.0013>; Paul T. McCartney, *Power and Progress: American National Identity, the War of 1898, and the Rise of American Imperialism* (Baton Rouge: Louisiana State University Press, 2006).

the same number of topics for each decade despite this, to make a comparison across time more valid. Additionally, some decades suffer from particularly poor OCR, either because they were digitised from a different medium (artefact or microfilm), using a different version of OCR software, or from an artefact that is harder to read (eg. Typeface, paper quality, level of preservation). Another possible reason for the low numbers of articles returned may be the keywords themselves: for example, Germany will be rare in newspapers before 1870, as it did not exist as a unified country and was split into (mainly) Prussia, Bavaria, Saxony and Baden-Württemberg.[112] However, using a consistent keyword, even if it results in less returns than desired in some periods, is preferable to going on a "Boolean fishing expedition".[113] Alternatively, it may point to there being little in the way of imperial or foreign news being reported in some titles in particular.

Despite the issues mentioned above, the aggregated topic model composition, scaled to 100% (figure 5.5), still shows the relative dominance of political reporting, which generally ran between half and a quarter of the coverage surrounding both the empire and foreign news, as shown by the graph below. Only in the first and the last decade does the empire overtake foreign news in a significant way. Additionally, there appears to be no correlation between the amounts of foreign news and large international political events, such as the Crimean War or the Franco-German War. However, there is a change visible in the nature of foreign news that is reported, in particular in its form. With the rise

[112] Georgi Verbeek, 'De wording van een Natie: Duitsland tijdens de lange 19e eeuw', in *Een Geschiedenis van Duisland: Sporen en Dwaalsporen van een Natie* (Leuven and The Hague: Acco, 2010), pp. 105–53.
[113] Underwood, pp. 64–65.

of the telegraph in the 1860s, the way newspapers gathered news significantly changed. Initially, correspondents employed by newspapers still wrote the copy the same way, only telegraphing it back to their editors instead of mailing it. This changed during the 1860s, and short headlines became more common – a development that was in large part fuelled by the rise of news agencies such as Associated Press and Reuters.[114] This is seen as one of the main reasons W.H. Russell, whose literary and extensive style of writing brought the Crimean War to life for many British readers, failed to repeat this success fifteen years later during the Franco-German conflict.[115]

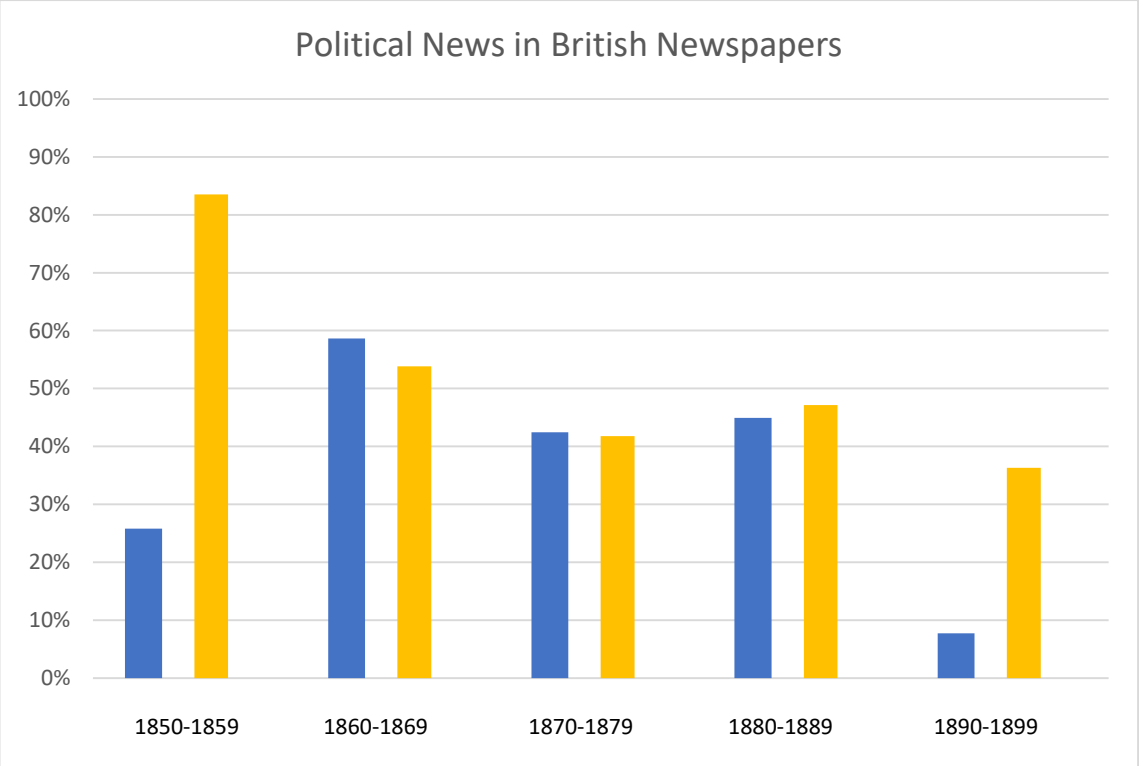


Figure 5.5 Political news in Foreign (blue) and Imperial (yellow) subsets. In order to compare these two subsets of different size, the size of the topics is scaled to be a percentage of subset.

[114] Standage, pp. 150–58.

[115] Markovits, pp. 25–29; Mitchel P. Roth and James Stuart Olson, *Historical Dictionary of War Journalism* (Greenwood Publishing Group, 1997), pp. 265–66.

The way the fraction of both imperial and foreign political reports change in the first and last decade are atypical, and suggest either a significant break in journalistic practices, or a glitch in the topic model. I consider it unlikely that these sharp declines of foreign and corresponding rises of imperial news are caused by the former. After all, for the 1850-59 subset we would actually expect a larger proportion of foreign political news: both France and Russia were involved with Britain in the Crimean war, which we know from literature was widely covered in the British press.[116] However, as this particular model was useful for other case studies, we simply have to live with its inaccuracies – no topic model will be perfect.

However, the nature of the foreign news in the topic models supports this view of rapid change in newspaper tastes. In the 1850-59 period, there is no telegraphic reporting and all foreign news is in longform, the 1860-69 subset contains both, and all subsequent subsets contain less than 1% longform foreign news articles. The only exception to this is if foreign countries are mentioned in Parliamentary debates and speeches. Imperial reporting, on the other hand, retains more longform news, which is reported on alongside the short telegraphic bulletins. In the 1890-1899 subset, long narrative imperial news articles still provide 30.5% of the uses of imperial locations.

Frustrations with the topic modelling continue when we attempt to model the *Huddersfield Chronicle*. This title has a specific issue that makes generating and interpreting topic models a challenging endeavour: the way the archive has chosen to handle article segmentation. This is a choice that the archival curator makes in

[116] Markovits.

their role as what James Mussel calls the editor of the archive.[117] The choices made by this editor, in this case, define what we can do, because they chose to define an 'article' in a specific way. In the case of the *Huddersfield Chronicle*, this has resulted in the column, for example 'Foreign News', being classed as the 'article', instead of slicing on the sub-headings of 'Asia', 'America', etc. Thus, within this archive-article, there are a collection of paper-articles, varying from a few lines on news received from India to a longer piece on relations between Britain and France. This kind of article is not unusual: however, for an unknown reason in the digitisation process segmentation is worse for the *Huddersfield Chronicle*.

For the purposes of understanding the imperial connection in the local press, this other material is 'noise' that we are not interested in. However, in the case of this paper, the signal-to-noise ratio is such that topics start to form along the lines of the other material. Because the document that contains the context of the keyword's use also contains multiple other texts that have nothing to do with it, the topic model will form on this other content, rather than on the imperial keyword being used. For example, the American civil war reports that surround the various imperial news items a few lines down get handled as a single document, which contradicts the basic assumption on which topic models rely: that each text has one dominant topic.

This is particularly frustrating, as the *Huddersfield Chronicle* has clearer OCR than most other papers, which would have made it an exceptional candidate for

[117] J. Mussell, *The Nineteenth-Century Press in the Digital Age* (Basingstoke: Palgrave Macmillan, 2012), pp. 117–21.

topic modelling, if it were not for the article segmentation issue. Additionally, contrasting with all the other papers, the *Huddersfield Chronicle* has the highest variance in the number of articles returned for each keyword of any paper in this case study, running from 3,409 for the 1850s to 15,717 in the 1890s. This means that while early topics are likely to be small and specific, with a particular risk for broken topics, for the later period there is enough material to work with. If the article segmentation does collaborate too, some very specific topics can form.

For example, the 1880s model contains one specific topic that directly links the political dimension of Britishness discussed by Ward with the empire. It includes a variety of tracts and calls to vote, including one that is very noticeably Conservative. This one is interesting as it explicitly calls for an expulsion of the liberal government based on its mismanagement of colonial questions. It makes most of these accusations before discussing the tax on beer or the income tax – items that would impact the voters directly.

> Remember the calamitous war and the dishonourable peace in the Transvaal; Remember the disaster at Majuba Hill, and its consequences; Remember the abandonment of Kandahar, the dismantled railway, and the loss of a defensible frontier in India; Remember that Lord Darby has gone far to estrange some of our most important colonies; Remember the bombing of Alexandria, the slaughter of Hioks Fasha's army, the massacres at Sinkat and Tokar — all the results of Liberal policy; Remember the bloodshed at Tel-el-Kabir, and in the two Soudan expeditions — all to no purpose. Remember Gordon.[118]

This article, and the topic it belongs to, clearly support the thesis that the empire was not only present in the lives of the British public, it was an active

[118] Anonymous, 'Facts for Electors to Remember', *Huddersfield Chronicle* (Huddersfield, 27 August 1885), p. 3, HUCE-1885-08-27-0003-024.

presence. In this case study, we have seen the empire appear in a variety of unexpected places within the local press. This shows that the empire, with all its global splendour, intruded into the everyday life of the citizens of an average English town, where it became part of its political discourse. A more covert invocation of the empire in politics occurs in 1880, when a writer advocates for "The political opponents of the government" to accept the annexation of Kandahar, as "the nation will not be exposed to war whenever a new minister comes to power."[119]

We thus have to conclude that while they occasionally produce interesting topics that prompt research questions worthwhile of entire theses, topic modelling this specific subset is not producing consistent results. For a variety of reasons, topic models do not provide a significant insight into the way political discourse in the British press facilitated imperial identities. Through a combination of relatively small datasets; issues with article segmentation and OCR; and limits to which, semantically, topic models can understand language, for two of the three decades we can't provide an accurate model. This case study thus forces the conclusion that on this subset, topic modelling is not the right approach to understand the relationship between foreign and imperial news.

Visualisation

Because of the failure of topic modelling, we have to deploy other tools to gain insight into the difference between reporting on imperial and foreign political affairs. This is where the visualisation tool developed in the previous chapter will

[119] 'The Retention of Candahar', *Southampton Herald* (Southampton, 14 September 1880), p. 2.

be tested. It aims to address two questions: were there differences in the space within a newspaper title where foreign and imperial news were covered? And are these differences consistent between papers? This requires a paper-by-paper approach, as the meaning of space varies between papers. For example, the second and third column of the sixth page in one paper may be dedicated to satirical jokes, while the same columns in another title may be for advertisements. In order to study the use of space of a newspaper's page, we need to gain some basic understanding of the underlying theory. Little has been done on the design principles of historical newspapers, thus we have to rely on modern scholarship, which is more plentiful.[120] For the purposes of this analysis, we will use the terms design and layout interchangeably. The key difference between these is that in layout one places a set of given building blocks on a page, while in design one also creates these blocks. This distinction is less relevant here, as both these processes took place in the construction of the newspaper page, undertaken by different people in the production process. However, as we are not interested in the actions of individuals but in the result of the entire process, thus design and layout may be used as synonyms.

The paper chosen to test the visualisation tool is *Reynold's Newspaper*, as it was among the most circulated papers for the working classes in the second half

[120] Kurvits, 'The Visual Form of Estonian Newspapers from 1806 to 1940 and the Appearance Spiral Model'; А. Киселев [Kiselev], 'История Оформления Русской Газеты (1702-1917 Гг.) [History of the Form of Russian Newspaper (1702-1917)]', in *The Visual Form of Estonian Newspapers from 1806 to 1940 and Teh Appearance Spriral Model*, by Roosmarii Kurvits, Nordicom Review, 29, 2008, pp. 335–52; Ven-Hwei Lo, Anna Paddon, and Hsiaomei Wu, 'Front Pages of Taiwan Daily Newspapers 1952–1996: How Ending Martial Law Influenced Publication Design', *Journalism & Mass Communication Quarterly*, 77.4 (2000), 880–97 <https://doi.org/10.1177/107769900007700410>; Johanna Schindler and Philipp Müller, 'Design Follows Politics? The Visualization of Political Orientation in Newspaper Page Layout', *Visual Communication*, 17.2 (2018), 141–61 <https://doi.org/10.1177/1470357217746812>. See for an introduction: Bob Franklin (ed.), *Pulling Newspapers Apart: Analysing Print Journalism* (London and New York: Routledge, 2008).

of the nineteenth century.[121] Being a weekly paper, *Reynold's* was cheap, with a launch price of 4 pence, later dropped to 2½ pence, putting it within the means of a working man. After the abolition of stamp duty in 1853, this price dropped even further to 1 penny. Its low price attracted many readers, with its claimed circulation in 1870 being 200,000. The paper appeared only on Sundays, further cementing its position as a working-class paper; Saturday afternoons and Sundays were the only days off for most labourers.[122] *Reynold's* offers a view into the media consumed by the working class, and its consistent layout helps with the interpretation of the results. It is also one of the more left-leaning papers and was noted for often sporting an opinion piece from a republican perspective as the lead article on the front page.[123] It was also noted as being opposed to imperial expansion for expansion's sake.[124]

Before exploring the meaning of the spaces occupied by articles, we first need to gain a more general understanding of *Reynolds's* pages as a space. Figure 5.6 shows two selected pages; from these we can see *Reynolds's* went through a redesign in 1885/86, changing from a six-column layout to a seven-column design. This was completely unexpected, as M. Shirley describes the paper as being laid out on "eight larger pages and eight columns". [125] A verification using the images themselves

---

[121] Virginia. S. Berridge, 'Popular Journalism and Working Class Attitudes 1854-1886 : A Study of Reynold's Newspaper, Lloyd's Weekly Newspaper and the Weekly Times.' (unpublished Ph.D., Birkbeck (University of London), 1976), pp. 101–4 <https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.449650> [accessed 22 March 2019].

[122] William J. Baker, 'The Making of a Working-Class Football Culture in Victorian England', *Journal of Social History*, 13.2 (1979), 241–51 (pp. 242–43).

[123] *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland*, ed. by Marysa Demoor and Laurel Brake (London: British Library, 2009), pp. 540–41.

[124] Antony Taylor, "'Some Little or Contemptible War upon Her Hands": Reynolds's Newspaper and the Empire', in *G.W.M. Reynolds: Nineteenth-Century Fiction, Politics, and the Press*, ed. by Anne Humphrey and Lois James (London: Routledge, 2017), pp. 98–120 (p. 118).

[125] Shirley, p. 540.

confirms the presence of six and later seven columns. Having more columns allowed the editor more options when composing the various articles into a coherent page and made more room for adverts, though at the cost of column width.

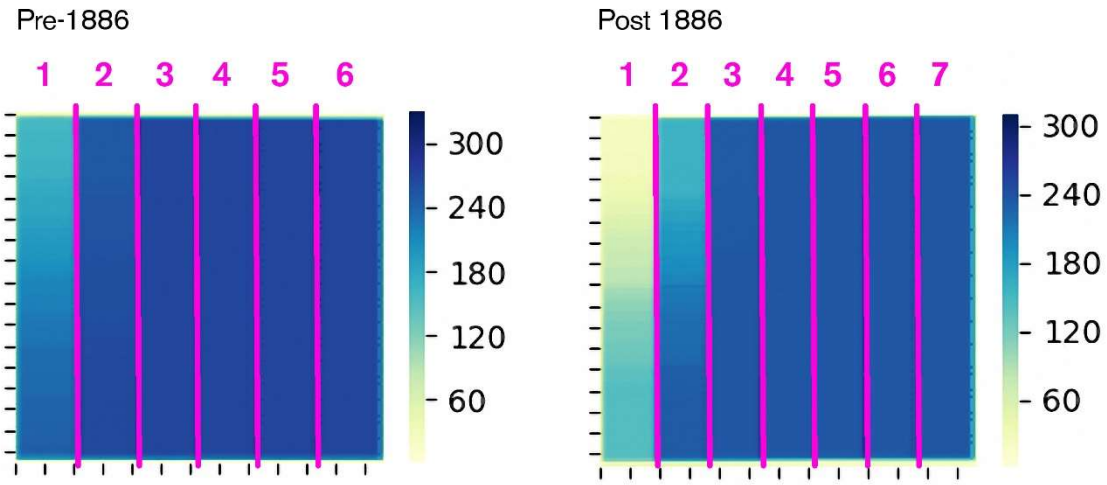


Figure 5.6 Columns in Reynolds Newspaper before and after the 1886 redesign. Only page 7 is shown. This illustrates the difference in number of columns observed by heatmapping.

During the period covered in this thesis, *Reynolds's* never switched to a columnless or modern layout. However, this does not mean it was without an underlying design philosophy: as observed by Barnhurst and Nerone, despite the appearance of disorder and immutability of newspaper design in the second half of the nineteenth century, rational and bureaucratic design elements did emerge. The centre and the periphery were made visual on the page by dynamic whitespacing, which had the most central and important content in airy, double-spaced lines on the top left and centre columns, with the density of the text increasing towards the bottom-right, where more peripheral content was placed.[126]

[126] Kevin G. Barnhurst and John Nerone, *The Form of News: A History* (Guilford Press, 2002), pp. 79–81.

However, whatever the editor intended the placement of content to be, the final say on the allocation of space was reserved for the foreman at the printers. Usually, his task required cutting back material or cramming in content wherever it would fit. These practices continued into the 1870s.[127] However, more recent newspaper scholars, such as D. Liddle, have argued that Victorian newspapers were very much in flux, with the information density, and by necessity, organisation, of their pages changing throughout the century. He states that the pages and genres within them only stabilised during the latter decades of the nineteenth century.[128]

Based on the article visualisations (Figure 5.7 and 5.8), we may conclude that *Reynolds Newspaper* did possess a consistent layout for the entirety of the period investigated. This conclusion derives from the clearly present clustering of articles for both keyword-selected subsets: if articles that contain the same keywords, and thus cover the same or similar topics, appear in similar places over a long period of time, we can safely consider there to be evidence for a 'rational and bureaucratic' design being imposed on the paper. The way these clusters shift shows that after 1886 the paper was redesigned in such a way that the political content, both foreign and imperial, occupied a very different space, barring one major exception. The most obvious section where the imperial and the foreign subsets overlap is in the first two columns of page 4 after 1886, which contains the densest concentration for both these families of article by far. With between 250 and 300 articles, these

[127] Barnhurst and John Nerone, *The Form of News*, p. 75.
[128] Dallas Liddle, 'Reflections on 20,000 Victorian Newspapers: "Distant Reading" The Times Using The Times Digital Archive', *Journal of Victorian Culture*, 17.2 (2012), 230–37 (pp. 231–34) <https://doi.org/10.1080/13555502.2012.683151>.

rows are hotbeds of occurrence. Yet these are not the only imperial spaces within the newspaper: from the front page to the last column, the empire is everywhere.

Their appearance becomes even more intriguing once the space is given proper context: these columns housed the very popular 'Notices to Correspondents' section of the paper. These kinds of sections have been theorised as "the principal forum for reader opinion", and the space for items that have "earned their legitimate place in the public debate".[129] Started by *Reynolds's* as a way to connect with his audience, readers could send questions, both on the mundane and the political to the newspaper's offices, which the editor (G.W.M. Reynold's himself until his death in 1879, subsequently his brother Edward and son William) would respond to.[130] 'Notices to Correspondents' helped create a community of readers and was an integral piece of the successful formula that saw it reach a claimed circulation of over 200,000. The change we observe in the placement of political content in general, but of foreign political content in particular, from adverts, economics and news sections to these discussion pages is suggestive of a deeper change in the way newspapers were read in the nineteenth century.

[129] John Richardson, 'Reader's Letters', in *Pulling Newspapers Apart*, by Bob Franklin (London and New York: Routledge, 2008), pp. 56–66 (pp. 56–57).
[130] Shirley.

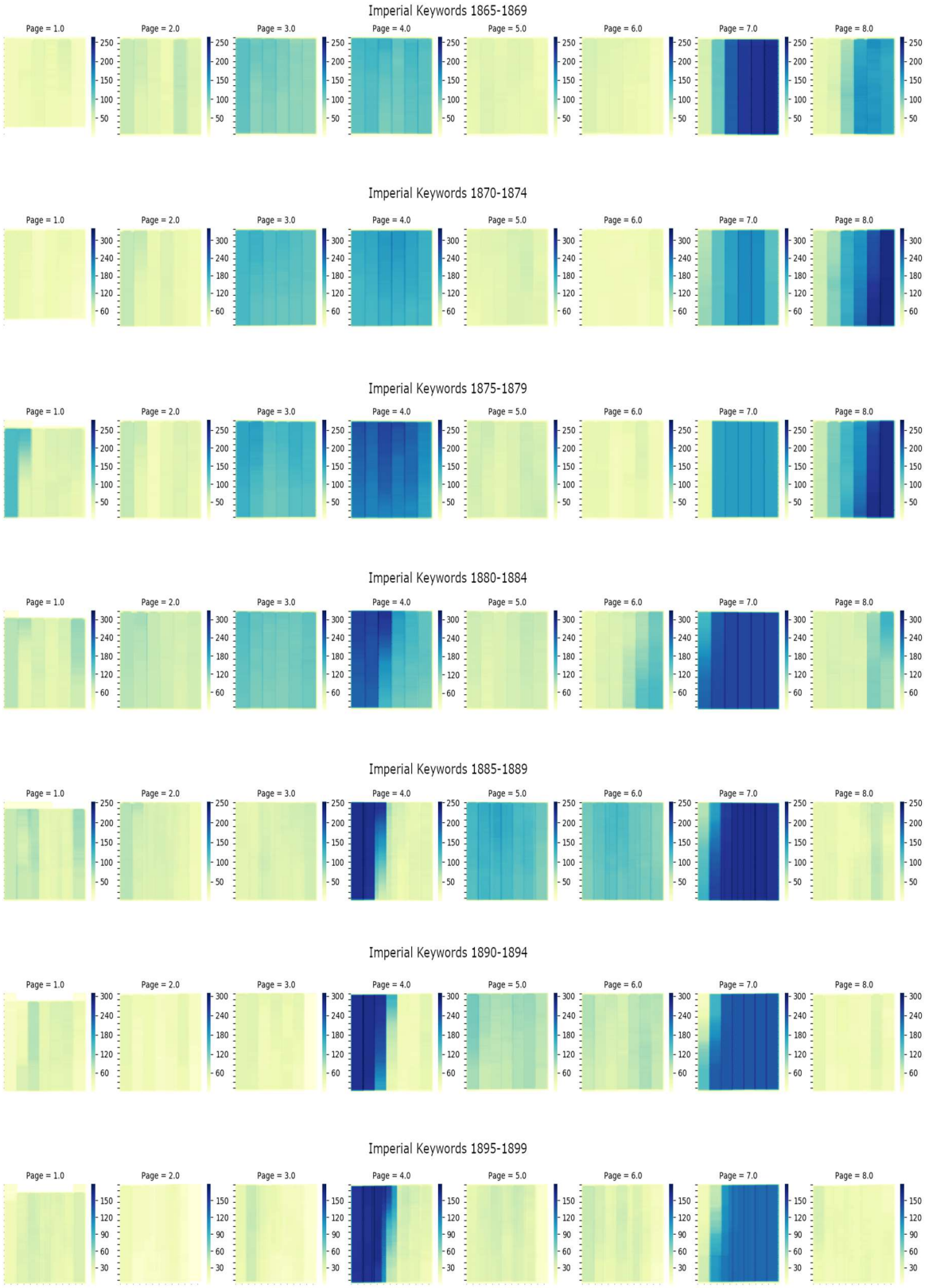


Figure 5.7 Imperial article placement in Reynolds Newspaper visualised in five-year intervals (January 1st of the first year to December 31st of the last). Keywords used: 'India', 'Canada', and 'Australia'. Note the prevalence of columns 1 and 2 on page 4 for the last three visualisations, representing the imperial debate in the 'Answers to Correspondents' columns.

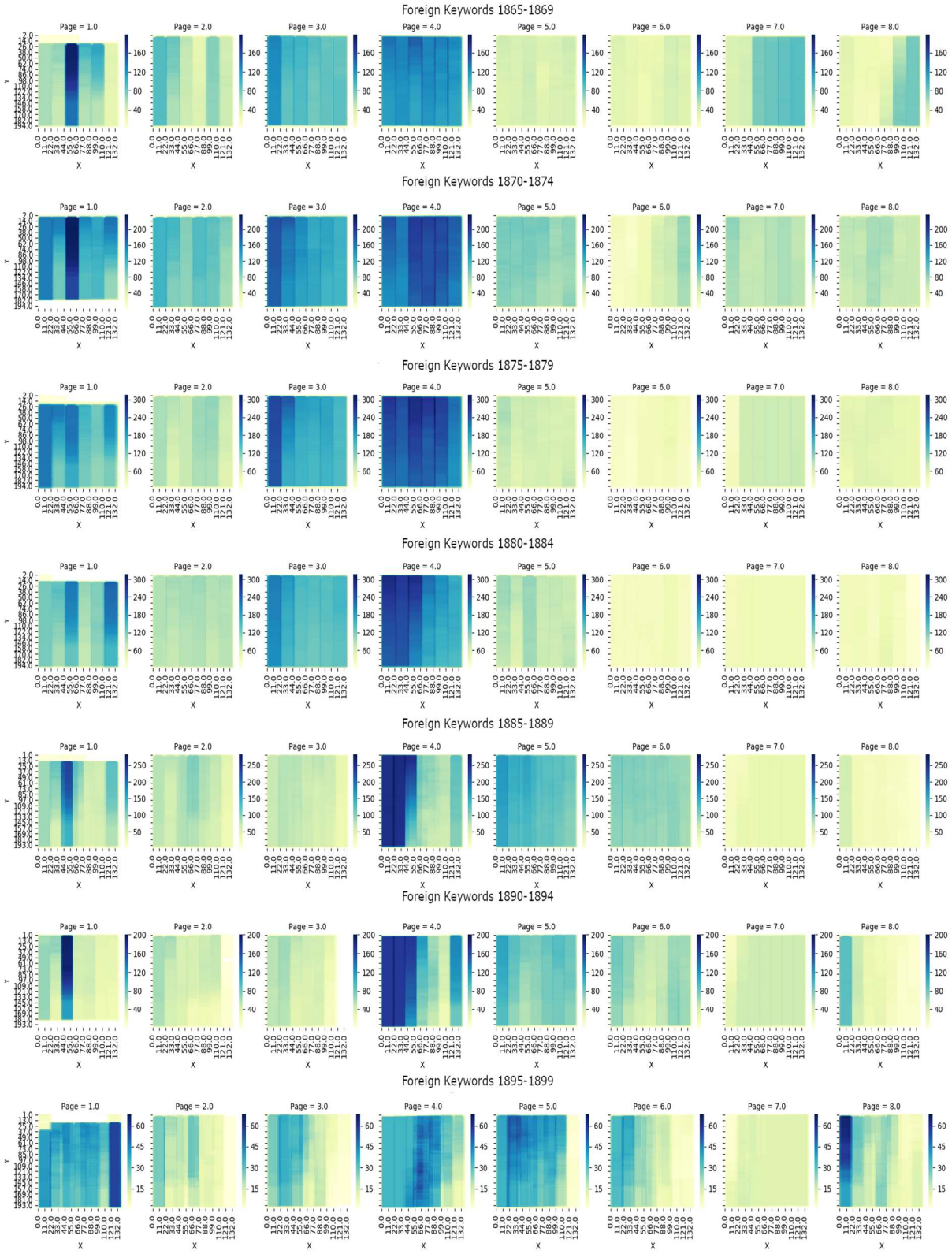


Figure 5.8 Foreign article placement in Reynolds Newspaper visualised in five-year intervals. Keywords used: 'France', 'Germany', and 'Russia'. Note the presence of foreign news on the front page compared to imperial keywords.

While newspapers offered a forum for public discussion since their inception, the rise of 'new journalism' saw these interactive practices reach new heights.[131] This heatmap seems to confirm those readings, as they show the newspaper's shift from being a vessel for news and information about the political developments of the world to being a platform for debating those politics. This would match McKenzie's reading of imperial identity, as it shows a large emotional investment by the average Britton into the empire. However, we could also read this as an example of the agenda-setting powers of the press: starting and fostering debate on topics in a way that forces political parties to respond.

The dominance of 'Notices to Correspondents' in the discourse of the empire and the foreign in *Reynold's* is remarkable; it outperforms its closest competitor, the advertisements on page seven, by a wide margin. But making note of this is as far as computer-supported research methods can take us: closer reading is needed to show the ways in which these places were referenced. It showed two contexts in which both the empire and the foreign Other were used. First, they may be invoked in direct response to a question from a reader. For example, in response to a query by 'W.G' in 1851: "Theatrical managers in France are made to pay one-tenth of the receipts to the support of the poor."[132] Or in October 1852, in an answer to A. Milner of Glasgow, when the response reads "If your health, age, character, &c., is such as to meet the view of the Government Emigration

[131] Joad Raymond, 'The Newspaper, Public Opinion, and the Public Sphere in the Seventeenth Century', *Prose Studies*, 21.2 (1998), 109–36 <https://doi.org/10.1080/01440359808586641>; Hannah Barker and Simon Burrows, *Press, Politics and the Public Sphere in Europe and North America, 1760-1820* (Cambridge: Cambridge University Press, 2002); Kate Jackson, *George Newnes and the New Journalism in Britain, 1880–1910: Culture and Profit* (Burlington: Ashgate, 2001), p. 76.
[132] [G.W.M. Reynolds], 'Notices to Correspondents', *Reynold's Newspaper* (London, 25 May 1851), p. 8.

Commissioners, a person of your calling might obtain a free or assisted passage to Australia."[133] In both cases, the original query obviously specified that they were inquiring into something foreign or imperial– in this case the payment into a social security system by French theatre owners or the ways for a 'man of certain calling' to make it to Australia. With the knowledge that these columns are the nexus for discussing foreign events in *Reynold's*, and realising their importance in the transmission of ideas about foreign policy, a more dedicated study would be able to perform a computational social science analysis on the 'Notices to Correspondents'. Such an investigation could mirror those that recently have relied on Google search queries to explore social issues such as (fears of) epidemics and (un)employment.[134] Only here we would be using the answers, not the questions.

While it is outside the scope of this thesis, such usage, however, does show that readers were interested in the places that were not commonly discussed in the rest of the paper. The second invocation of imperial and foreign places in the editor's responses is as a yardstick by which a certain factoid or measurement is to be taken. In these cases, the enquirer is asking after foreign practices to compare them with those within Britain. These are much more difficult to identify, as without the original query, they can look much like the simple queries, and a degree of close reading is required to find them. One example is the response to a question on the use of colonial troops by the French by 'M.E.' of Wye: "The Zauaves [sic] are natives of the French provinces of Algiers, disciplined and exercised by French

[133] [G.W.M. Reynolds], 'Notices to Correspondents', *Reynold's Newspaper* (London, 31 October 1852), p. 8.
[134] Hyunyoung Choi and Hal R. Varian, *Predicting the Present with Google Trends* (Rochester, NY: Social Science Research Network, 1 June 2012) <https://doi.org/10.1111/j.1475-4932.2012.00809.x>; Claudio Cioffi-Revilla, *Introduction to Computational Social Science: Principles and Applications*, Texts in Computer Science, 2nd edn (Springer International Publishing, 2017) <https://doi.org/10.1007/978-3-319-50131-4>.

officers, and now forming part of the French contingent employed in the Crimea. They hold exactly the same relation to the French army as the Sepoys in India have to the regular British troops."[135] In this case, the editor recognised that by making a comparison between an unknown foreign entity and a known imperial one, the reader would better understand the former. Another example in which the foreign is used as a comparator for Britain is a response from 1852, when discussing the subject of prostitution. "Prostitution is so far encouraged in France that it is made a licenced trade by the government. -No such custom as that you mention is tolerated by police authorities abroad."[136] This response shows a rare glimpse of the editor engaging with a preconceived notion in the mind of the correspondent. This response fits in the context of the 1850s marking the beginning of 'purity politics' and a political and moral campaign of repression that would cumulate in the 1886 repeal of the Contagious Diseases Act and 1885 Criminal Law Amendment Act. Here, we see a foreign example being used to contextualise an emergent British debate.[137]

The placement of political imperial news suggests that it is placed alongside or amongst articles covering national matters. The articles that are keyword-selected with imperial keywords, and also occupy political spaces in the newspaper do so mainly in the parliamentary columns. Here, the empire is discussed as an

[135] [G.W.M. Reynolds], 'Notices to Correspondents', *Reynold's Newspaper* (London, 12 November 1854), p. 8.

[136] [G.W.M. Reynolds], 'Notices to Correspondents', *Reynold's Newspaper* (London, 10 March 1852), p. 8.

[137] Philip Howell, David Beckingham, and Francesca Moore, 'Managed Zones for Sex Workers in Liverpool: Contemporary Proposals, Victorian Parallels', *Transactions of the Institute of British Geographers*, 33.2 (2008), 233–50 (p. 236) <https://doi.org/10.1111/j.1475-5661.2008.00292.x>. For more context see: Frank Mort, *Dangerous Sexualities: Medico-Moral Politics in England Since 1830*, 2nd edn (London and New York: Routledge, 2000).

integral part of British political life. However, news from parts of the empire, relating to (political) events that take place there are not reported separately. This suggests the empire does hold a special position in British political discourse, and in the political identities it imparts on its citizens. It is not merely an overseas place, but one that, because of the power Britain holds over it, is a space that the British social, economic and political lives intersect with on a regular basis. While the foreign only appears when it is relevant, the empire is on every page, as it is always relevant to Britain's political debates. This relevance is not always solely a geopolitical one; as the extract from the *Huddersfield Chronicle* shows, the far reaches of the empire may very well be invoked for extremely local political capital.

In conclusion this case study has further highlighted the limitations of topic modelling tools for historical newspaper research. While they are powerful when used to classify articles, they are unable to comprehend the text that they sort, and are therefore susceptible to issues with the source material. Article segmentation is especially important to generating good outputs. However, when they are employed in an exploratory manner to point to various contexts in which the keywords that were used to generate the subset appeared, they are still valuable. They may still be used to verify high-level assumptions made about the corpus; in this case, the dominance of political news predicted by historiography was verified by the topic models. Furthermore, this case study has shown the usefulness of visualisation tools for newspaper space, which opens a new dimension of the source material to be explored. Its use has allowed this project to shed new light on existing discussions in newspaper history, such as the distribution of content throughout a title.

Case Study Six: Leisure and Entertainment

This final case study will explore the way imperial ideology was transmitted in the British leisure press. This category of newspaper article includes both sections of dedicated publications, such as seaside resort newspapers, and columns of the mainstream press, such as horse racing returns in the *Liverpool Mercury* or cricket scores in the *Daily Telegraph*. This case study's choice of subject was directly informed by the fifth aspect Ward identifies as part of British national identity: leisure. By consuming a shared body of media, enjoying shared sporting fixtures and co-inhabiting the same holiday spaces, the British could share an experience of free time. This case study looks at the imperial intrusion on this experience: is there evidence to suggest the leisurely pursuits of Britons were shaping their feelings of belonging to the empire? Did Victorian popular culture propagate imperialism, as McKenzie suggests? As in the previous case study, a comparison will be made between Imperial and foreign markers of place. Three forms of leisure are particularly relevant: seaside resort vacations, theatre, and sports.

The first, vacationing in seaside resorts, is best investigated using a local newspaper. The paper chosen for this is the *Hampshire Post and Telegraph*, which is the oldest of the titles used in this chapter. It was founded in 1799 in Portsmouth and took the name *Hampshire Telegraph* in 1803. It was amongst the first to report on the Battle of Trafalgar – helped by the maritime connection of the city. It published on a weekly basis. Its naval roots are very noticeable in the topic models it generates: of all the papers it continues reporting on the arrival and departure of packet steamers longest. After they disappear from the topic models of the *Preston Chronicle* and the *Huddersfield Chronicle* in the 1870s and 1860s respectively, the

Hampshire Telegraph keeps publishing them for the entire century – admittedly in smaller volumes.

The most intriguing aspect of the empire that shows in the topic models for the *Hampshire Telegraph* is the consistently growing presence of the column 'Southsea Visitors'. To understand this column, we need to understand the context which produced it: the Victorian leisure industry. This industry developed significantly in the latter half of the century, especially after the 1877 Bank Holiday Act. Day trips and longer holidays to the seaside were becoming commonplace, and towns like Brighton and Blackpool were amongst the fastest growing in the country, offering city-folk a literal breath of fresh air.[138] The town of Southsea, just outside Portsmouth at the tip of the Portsmouth peninsula was another such a city. Portsmouth had been connected to London by rail in 1847, which allowed travellers to make a day trip to the city, where a horse-drawn bus service connected with the train to take passengers to Southsea.[139] These would be more affluent middle-class urbanites: the centre for working-class seaside tourism was Margate.[140]

The *Hampshire Telegraph* published a weekly list of people that had come to visit the seaside town, which hotels they were staying in, how long they were staying, and where they had come from. This information was gathered by leaving books in hotels and guesthouses where people could submit their details. For the

[138] Cannadine, *Victorious Century: The United Kingdom, 1800-1906*, pp. 512–14.
[139] Peter Gould, 'Portsmouth Corporation Transport: 1901-1986', *Portsmouth Corporation Transport, 1901-1986*, 1999 <http://www.petergould.co.uk/local_transport_history/fleetlists/portsmouth1.htm> [accessed 18 June 2019].
[140] Allan Brodie, 'The Brown Family Adventure – Seaside Holidays in Kent in the Mid-Nineteenth Century', *Journal of Tourism History*, 5.1 (2013), 1–24 (pp. 18–19) <https://doi.org/10.1080/1755182X.2012.758671>.

purposes of this case study, only one of these characteristics is notable: origin of the tourist. People from India and Australia, but mainly Canada, staying for longer than the average tourist, feature in good numbers on these lists. This shows the intermixing of colonials and Britons in the most voluntary space possible: on holiday, sharing the beach during the day and the music hall at night.

These imperial visitors stayed longer than average. Colonel Lyons, from Canada, is recorded on at least four such lists, staying in the Southsea Beach Mansion Hotel between the 30th of May and the 25th of July 1868, longer than any other guest recorded.[141] In some cases they even stayed together in the same apartment. On the 7th of August 1882, the guesthouse at no. 8 Southsea Terrace was recorded as: "-No. 8: Mrs. and Miss Cox, from Norwood: Dr. and Miss Salmon, from India; Major and Mrs. Tigott and family."[142] In this case, imperials and Brits mingled under the same roof in a holiday setting.

This only leaves the puzzle as to why these lists were published in the first place. The *Hampshire Telegraph* is not alone in publishing these, as many other seaside papers had them. The *Ilfracombe Intelligencer* published them on their front page, while some resort newspapers even existed almost exclusively for this purpose.[143] While these lists of visitors have been used as sources for local history, there is no consensus on their original purpose – although the wider social aspect

[141] 'The Southsea Visitors' List', *Hampshire Post and Telegraph* (Portsmouth, 30 May 1868), p. 6; 'The Southsea Visitors' List', *Hampshire Post and Telegraph* (Portsmouth, 25 July 1868), p. 6.
[142] 'Southsea Visitors', *Hampshire Post and Telegraph* (Portsmouth, 7 August 1882), p. 3.
[143] Andrew J. H. Jackson, 'Provincial Newspapers and the Development of Local Communities: The Creation of a Seaside Resort Newspaper for Ilfracombe, Devon, 1860–1', *Family & Community History*, 13.2 (2010), 101–13 (pp. 104, 110) <https://doi.org/10.1179/146311810X12851639314110>.

of Victorian seaside resorts has been the subject of intense study.[144] The main theory is that these lists played an important role in the social fabric of the seaside community, allowing residents to take note if acquaintances were in town on holiday, and allow friends to find one another.

The inclusion of the marital status of some of the submissions also suggests that these arrival lists played another function. In the contemporary satirical piece *London out of Town*, written by John Leighton, the seaside is explored as a place of courting and romance, though also a place in which unscrupulous men can take advantage of a young lady's feelings – it is a comedy after all.[145] It is thus entirely reasonable that these lists were also a way for people to make it known that they were 'available' to those of a similar social class. The topic models showed that on holiday by the seaside, the (middle-class) Victorian holidaymaker would encounter the empire not just as a flag on a page, but as an actual person with whom personal, individual connections to the empire could be formed at a dinner party or dance hall.

Holidays were rare events that a typical family would only indulge in once a year at most, but a visit to the theatre would be within the means of a household much more often. Based on the topic modelling for both the keyword sets of imperial and foreign locations however, there is little that can be said: no topic appears with these keywords that contains theatre adverts. They do appear within some of the foreign subset, but as almost accidental bit topics with just three or

[144] J.K. Walton, *The English Seaside Resort: A Social History, 1750-1914* (Leicester: Leicester University Press, 1983).
[145] Brodie.

four articles to them. This absence is itself telling, as it suggests that in the narrative surrounding theatre and its traveling performers, there were little to no connections with either the empire or other countries. One of the sole exceptions is notices like the following from *Reynolds Newspaper*, commenting on the artistic fashions of other countries:

> It has been remarked that the ballet has lost its attractions for our opera-going public, which may be supposed to represent the highest class of society In England. From the opera houses it has come down to the theatres frequented by the middle classes, and is even more appreciated at the music-halls than at the theatres. A similar outage of taste would seem to prevail in Germany, if we may accept the report that the Duke of Saxe-Coburg-Gotha [nb. Alfred, second son of Queen Victoria], who is his own opera manager, has lately disbanded his Corps de Balet (sic).[146]

However, one of McKenzie's main arguments is for the presence of empire in exactly the music halls and theatres, while our selection of imperial keywords is not finding any references to them.

Mindful of the potential distorting effect of keyword choice, the topic models were generated again using articles gathered by the keywords 'empire' and 'imperial'. After this change, theatre topics started appearing fast, forming away from the political commentaries on French and Russian empires. Yet this, too, is of limited use. While the appearance of this topic proves that between 4 and 9% of the uses of the word 'empire' in the British press related to the names of theatres, the presence of these theatres is a known entity. We are thus not using our tools to support historical research, but are simply engaging in a computer-assisted fishing expedition. A much more detailed study of the plays performed there using

[146] 'Foreign Intelligence', *Reynold's Newspaper* (London, 17 January 1869), p. 8.

these topics would be needed to make a claim on imperial penetration of the theatre landscape. Again, we reach the limits of where topic models can bring historical research. While it can gesture towards interesting avenues of research, it cannot replace historical methods.

Sports are more easily detected via topic models, as they contain very distinctive but repeated language. The words 'cricket', 'gelding' and 'derby' have little other uses in journalism. On the imperial subset, there are no sports topics that form with any kind of consistency, and when they do appear, they fold into more miscellaneous topics. This itself was a major surprise, as sports and imperialism have been extensively linked in historiography. J.A. Mangan considered the propagation of games from British public schools to the empire as "a dissemination of the moralistic ideology of athleticism, empire and imperial control".[147] Guttmann takes an entire chapter to describe the introduction of cricket into different British colonies, and discusses the way these games were adopted by the population as acts of acquiescence, as happened in India.[148]

Furthermore, we know that British sports teams made journeys to parts of the empire to tour and play matches, and that these were extensively covered in colonial newspapers. For example, the All England Eleven Cricket team toured Australia in 1861-62, an event which Frost argues was highly significant in the formation of Australian national identity.[149] International Test matches have been

[147] James Anthony Mangan, *The Games Ethic and Imperialism Aspects of the Diffusion of an Ideal* (London; Portland, Ore.: F. Cass, 1986).
[148] A. Guttmann, *Games and Empires: Modern Sports and Cultural Imperialism* (New York: Columbia University Press, 1994), pp. 15–40.
[149] W. Frost, 'Heritage, Nationalism, Identity: The 1861-62 England Cricket Tour of Australia', *The International Journal of the History of Sport*, 19.4 (2002), 55–69 <https://doi.org/10.1080/714001786>.

played between Australia and England since 1872, playing each other every two years for 'the Ashes'.[150] Yet only when taking just the articles found using the keyword 'Australia' and using these to form a topic model covering the entire fifty years and 21,000 articles, does cricket emerge as a separate topic. Even then, it represents only 0.004% of the coverage of Australia in the five newspapers used. This suggests that, for this period and based on these sources at least, cricket matches against Australia were not hugely significant events for British readers. This is surprising given the importance in literature given to the British-Australian Cricket rivalry, which considers the sporting rivalry shaping both Australian and British identity.[151]

We do need to place some caveats at this finding. First, it may be that most of the discourse around these matches took place in dedicated sports newspapers, while the regular titles in our dataset only reported on the matches infrequently, thus creating an impression of absence. To test this, we would need to run these topic models on a collection of digitised sports newspapers, but these are not available in the dataset used by this project. A way to see to what extent the Australian rivalry exists in these sources could be to run a keyword search for

[150] Rex Alston, 'Sports: Games of Bat and Ball: Cricket', *Encylopedia Britannica* (Chicago & London: University of Chicago Press, 1985), 130–36 (p. 133).

[151] David Frith, *The Golden Age of Cricket, 1890-1914* (Guildford: Lutterworth Press : Richard Smart Pub., 1978); David Frith, *England versus Australia a Illustrated History of Every Test Match since 1877*, 12th edn (London: Viking, 2007); Joseph Maguire, 'Globalisation, Sport And National Identities: "The Empires Strike Back"?', *Loisir et Société / Society and Leisure*, 16.2 (1993), 293–321 <https://doi.org/10.1080/07053436.1993.10715455>; Dominic Malcolm, 'Malign or Benign? English National Identities and Cricket', *Sport in Society*, 12.4–5 (2009), 613–28 <https://doi.org/10.1080/17430430802702897>; Dominic Malcolm, *Globalizing Cricket: Englishness, Empire and Identity*, Globalizing Sports Studies (London and New York: Bloomsbury Academic, 2013); Dominic Malcolm and Philippa Velija, 'Cricket: The Quintessential English Game?', in *Sport and English National Identity in a 'Disunited Kingdom'*, ed. by Tom Gibbons and Dominic Malcolm (London: Routledge, 2017), pp. 18–33; Ric Sissons and Brian Stoddart, *Cricket and Empire: The 1932-33 Bodyline Tour of Australia*, Routledge Library Editions: Sports Studies, 2nd edn (London: Routledge, 2014).

'cricket', and experiment to determine the size of an 'ashes' topic within that subset. Secondly, it may be that the mentions of Australia before 1872, in a non-sporting context, mask the cricket reports. But if this is the case, it shows exactly that the average British newspaper reader encountered Australia in significantly more contexts than cricket, far before matches began. Alternatively, it is possible that the wrong keyword was used, and that other forms of the word, such as 'Australian' or 'Australians' might have found more use in this context.

The situation appears very different in the foreign subset. Unfortunately, a closer reading shows that the keyword curse strikes again; what seemed to be very interesting appearances of cricket match reports in an international context turned out to be, on closer reading, a pair of Huddersfield brothers named W. and T. French who bowled for the local club for about two decades. However, there are multiple sports topics in this subset, and after cricket, the main sports topic concerns horse racing. Many of these are race reports, as well as betting odds, race calendars, and adverts for dedicated racing magazines with titles such as the *French and English Sportsman*, *French and English Turf Chronicle* and the *Continental Turf Gazette*. The topic model here prompts an interesting historical question: how interlinked were continental and British sports, how did this come to be, and what role did these sports newspapers play?

The topic first appears in the 1870s subset and persists in every subsequent decade, but does its presence signify an international dimension to British sports? Close reading of the articles suggest so: the race calendars cover the main English races as well as those in France and Germany, betting odds are given for horses in

races both abroad and in Britain, and several of these papers are listed as having offices in both France and London. However, closer investigation into the literature around horse-racing in the nineteenth century paints a different picture, and yet again drives home the importance of contextual knowledge when interpreting topic models. Before the 1874 Betting Act came into force, it had been common practice for tipsters and journalists to work together in making money from horse racing. In return for paying several times the normal rate for adverts, tipsters left blank spaces in their notices that the journalist would fill in with the name of the horse that actually won the race before the paper was printed, giving the tipster an air of infallibility. When this behaviour was prohibited by law, the tipsters (who were often side-lining as bookmakers) continued their practices from the safety of the continent. They began to publish their own newspapers from abroad, taking advantage of the telegraph and railway connections that joined Britain with France. They were then free to advertise these papers with a 'taster' in regular British newspapers, offering to send them to any address in the United Kingdom. When the French government tightened its laws in 1891, they moved their offices to The Netherlands and continued operations as they had before.[152]

These newspapers were aimed at all social classes, and their adverts are thus found in all newspapers in the corpus. Gambling on horse races was a popular pastime for members of working-, middle-, and upper class, especially as it was only lightly regulated and the chance for victims of fraud to recoup losses were

[152] James Lambie, *The Story of Your Life: A History of the Sporting Life Newspaper (1859-1998)* (Troubador Publishing Ltd, 2010), pp. 282–83.

slim.[153] Extensive protections for gamblers were seen as not only impinging on freedom of the press or personal liberty, but as gambling was a moral failure, people who lost it all only had themselves to blame.[154] However, while these papers are an excellent example of the world becoming ever more interconnected during the nineteenth century, they can't be used to anchor statements about British national or imperial identity in relation to sports news. The topic models offer no evidence that suggests that imperial sports matches played an important role in the formation of a British imperial identity, but may have been important to the national identity of the colony against which the game was played.

The results of this case study's topic modelling and visualisation suggest the impact of the empire on Victorian leisure culture was relatively minor, and that we cannot consider that it had an impact on the formation or propagation of an imperial identity based on the topic models generated. However, this should not be taken to mean that there was none, only that it was not detectable by the methods developed in this thesis. The models do show very interesting and intriguing topics. The seaside visitor lists, the foreign horse-racing newspapers and the unimportance of 'the Ashes' in the press all raise new and pertinent questions. Can we reconstruct demographics of tourism using these lists? What would the topic model of one of these imported betting newspapers look like, and does its composition change as it is forced to move counties? When did the public interest in the sports rivalry with Australia grow, and does the growing independence of

[153] Mike Huggins, 'Culture, Class and Respectability: Racing and the English Middle Classes in the Nineteenth Century', *The International Journal of the History of Sport*, 11.1 (1994), 19–41 (pp. 20–21) <https://doi.org/10.1080/09523369408713845>.
[154] Lambie, pp. 283–84.

the territory have anything to do with this? All these questions deserve a full historical study that this thesis is not suited to provide.

The value of topic models is that, even if they do not answer research questions, they prompt new ones that might otherwise have been overlooked. The importance of distinct categories such as visitor's lists, or continental gambling papers invites more research. Similarly, the lack of cricket news from Australia prompts a deeper investigation on the longstanding orthodoxy that Imperial Cricket competition formed an important part in British Identity – or at least how these matches were experienced, if not through newspapers.

Conclusion

The case studies in this chapter have addressed the methodological questions on the practical use of topic modelling and spatial visualisation tools posed by this project, as well as the historiographic questions on using these tools in conjunction with theory. Methodologically, it has shown that both topic modelling and spatial visualisation are useful methods for researching newspaper sources for traces of identity, as long as they are used within their respective limits. Spatial visualisation, in particular, was found to have significant potential as a research tool. The historiographic contribution of the thesis consists of contributing additional support to the Manchester School's thesis of popular imperialism, showing that Billig's banal nationalism holds when studying imperialism. Additionally, it has shown that three of Paul Ward's facets of British national identity after 1870 are applicable to imperialism from 1850 onwards, showing how digital tools can test

the applicability of historiographical theoretical frameworks outside the parameters they were initially designed for.

Over the course of the five case studies, it has become clear that topic models, while useful tools, have some hard limits. The case studies have established that topic modelling can be used to answer questions about the spread and size of certain categories of material, i.e. the topics, relative to the size of the corpus. These topics are limited by the process that produces them to only appearing around linguistically distinct texts; if a concept is only ever discussed in similar language as other concepts, it will not produce a topic. At their best, these topics can be extremely specific, but topic model performance is intimately connected to OCR quality, segmentation of the archive, and the linguistic distinctiveness of the sought category. Because these aspects are not within the control of the researcher, they are not easily measured, and can have a significant impact on the shape of a topic model, which means that they cannot be relied on as absolute evidence, but must be used with epistemological scepticism. However, even if the result of a topic model is uncertain, they can still be used as a prompt for further research by presenting linguistically atypical texts as part of their topic formation process. Where they contradict established literature, they offer an entry point for further study.

While the visualisation tool was only applied to one of the case studies, it offers much firmer foundations on which to build historical arguments than topic modelling. There are fewer factors that influence the accuracy of its result, which

gives it significantly more 'saying power'.[155] However, it is very situational, as the subset needs to be generated entirely from a keyword search. Therefore, it requires a situation where one or more keywords are capable of covering the entire concept that is being searched. If this can be done, however, visualisation allows for a deep overview of the space in which a keyword was used, which can lead to insights into reader engagement or editorial practice. However, the factor that limits their deployment is currently the computational requirement, which makes using these visualisations time-consuming and limits the amount of articles that can be drawn simultaneously.

Both of these tools were found to be extremely potent in providing a researcher with the context in which the articles existed while they were actively being read. Archives have the tendency to define their collections by the smallest unit of meaning they hold, which in the case of newspapers is the article, while losing sight of the ecosystem of other articles and papers with which it competed for the reader's attention. Topic models recreate the textual and linguistic context in which an article existed; visualisation does so for the spatial context. This spatial context is especially interesting, as it allows us to understand the way a regular reader would have experienced an article, knowing what had occupied its space for the preceding issues.

Yet the main conclusion on the use of these tools as means of supporting historical research has to be this: while they are powerful and useful additions to

[155] 'Zeggingskracht' in Dutch is a concept covering both expressiveness and the strength or certainty with which a statement is to be taken.

the historian's toolbox, they will always need a historian who manually analyses the results and interrogates anomalies. Therefore, topic modelling and article placement visualisation can only be used as an addition to the historian's arsenal, not as a substitution for the methods that have served for centuries.

Additionally, topic models offer the ability to measure the relative rise and fall of journalistic genres, if these genres align with topics. Such alignment is common, due to each genre having its own distinct vocabulary: a police court report will use different language than a parliamentary speech. This identification and measurement has the potential to be of great use to scholars of the history of the press itself. It could allow for a charting of the development of different genres, by producing a topic model of different newspaper tiles on a yearly basis and tracking the increase and decrease of the size of these topics in the same way these case studies analysed their topic sizes. Alternatively, topic models may be used for the comparison between titles that were considered 'new journalism' and 'traditional', in order to test the differences in writing style and subject choice that is to be expected based on scholarship.

As part of its methodological investigations, these case studies have not only enabled conclusions about the tools used, but also facilitated the debate on historiography, both on individual case studies and on a more theoretical level. In its first case study it established an all-encompassing overview of the composition of this archive, using topic modelling on a random sample. The second case study verified the longstanding connection made in literature between the empire and economic benefit for Britain, and showed that this connection was relevant for

both urban and rural readers. In the third case study it explores the empire through family notices. This case study shows there existed a much deeper and personal connection to the imperial project in the press than had been theorised before. The initial findings suggest that these family notices may be useful texts to use for investigations of ways the empire allowed a renegotiation of gender roles. In the fourth case study, it found that references to the army in an imperial context can be very explicit, forming topics around campaigns and battles. The navy, on the other hand, was dominant in advertising, especially after the 1870s. Both of these were also extensively reported on in everyday and banal movement lists, which would have kept relatives at home updated on the imperial adventures of 'their boys'. Additionally the navy was a particular tool for advertisers to signal a patriotic and imperial virtue. Finally, the leisure and entertainment case study highlighted some interesting initial findings. It discovered the visitor's lists of seaside newspapers to be a valuable source for exploring the intersection of imperial and national spheres. It also found that, contrary to the available literature, the importance of cricket in the identity relationship between Britain and Australia was negligible, at least in the newspaper press. This observation calls for more research to see if this holds when taking sports periodicals into account.

On a more overarching level, this chapter has shown there exited a widespread network of banal imperial flags throughout the British press. In market news and advertising, in military movement lists and family notices, in Parliamentary debate and seaside visitor's lists, every aspect of the newspaper press participated. Many of these are uses that can be understood within Billig's theories on the creation and spread of (national) identities through banal mentions. This

thesis has shown that such a banal imperialism was prevalent throughout the newspaper press between 1850 and 1900. These findings support the reading of MacKenzie and the Manchester School that imperialism was widespread in British society. These case studies have also established that these imperial elements were present in 1850, earlier than dates suggested by several other researchers such as Ward (1870), Summerville (1870), Springhall (1880) and MacKenzie (1880).

This chapter has also established that certain facets of Ward's schema for Britishness after 1870 are also applicable to imperialism before 1870. It established the gendered dimension of imperial family notices, and found a possible political facet in the way the press reported on imperial compared to foreign policy. Political columns presented the empire as integrated part of British life and the British political system; by contrast foreign political debates were placed explicitly separate in their own space. It also hinted towards national political tribalism gaining an imperial connotation, especially surrounding elections. The leisure facet gives a conflicting result: on the one hand the seaside visitor's lists present a genuine imperial space, on the other hand everyday sports events remain local and national. More work is required to definitively prove either case.

The spatial dimension of the everyday and banal encounters the Victorian reader had with the empire is telling. Prior researchers were unable to comprehend this dimension, as it was hidden behind archival interfaces and inaccessible metadata. However, the visualisation tool this thesis has developed allows the exploration of this facet of the nineteenth-century press, and shows the empire not as simply a space on a page, or a single page. Rather, the empire is spread across

the entire newspaper, with virtually no page left untouched. Some hotspots were expected based on scholarship and prior topic modelling. Examples of the topic models aligning with prior research include the economics section with its coverage of overseas markets being a topic of significant size, and the prevalence of the military stations list and the importance of the armed forces in forging imperial bonds. Other topics were much more unexpected and less in alignment with prior scholarship, such as the importance of family notices and the relative absence of Cricket in press coverage of Australia. Yet the visualisation tool is at its most powerful when used in a comparative analysis. When looking at the placement of imperial articles compared to British or foreign ones, the imperial and British articles collocate, while foreign news has its own, separate space. This pattern of spatial appearance is in line with Billig's theory of banal nationalism, as the empire is everywhere, just as the 'British' is, while the foreign other is confined to just a handful of places.

Conclusion

The moment has come to down the chisel, sweep up the workspace and wheel out the finished sculpture into the view of the public. This particular project was challenging, as it took place at an intersection of schools and styles, the material it worked on had an inherent roughness to it that it could not remove, and the tools it built to work with occasionally failed to perform as intended. Yet the core ideas that created it are present. For this thesis, these were four key questions. First, how can digital tools be integrated in the body of traditional historical research methods? What models should this follow and what is the role of matters such as source criticism and tool building? Second, how do historians sensibly integrate topic modelling into historical practice? Third, how can spatial visualisation of articles placement be used for historical research? And finally, can these digital tools make a contribution to historiographical knowledge?

In its first chapter, this thesis addressed the historiographical and methodological context in which it operates. This context has to be broad by necessity, as this project intersects three fields, History, Computer Science and the Digital Humanities, and has to allow readers from any one discipline to follow along with the project's contributions in all three. It did this by extending the literature review of a regular thesis, covering the fundamentals needed to understand the goals and challenges of this project. This approach is the solution that requires the least fundamental changes to the current historical research process, while still allowing for transparent and grounded tool criticism. In order

to perform this vital criticism, we need to both have a shared vocabulary, so that designer, programmer, and user can speak to each other; we also need to develop an understanding of the context in which tool design decisions were made, so that we can honestly critique them.

The thesis discussed the history of the *British Library 19th Century Newspapers I and II* archive in its second chapter. By doing so, it addressed questions of source criticism and laid the groundwork for exploring the advantages of building tools. As each archive is unique, the tools that seek to use it also have to have a level of unicity, at least if they interface directly with the archive. In order to construct tools that fulfil this requirement, and to critique such tools, it is essential to have a good understanding of the archive's history. This thesis has provided this overview, focussing on those aspects of the archive that influence this research project: the creation of the newspapers in the context of the nineteenth-century press, the troubled and occasionally chaotic history of conservation, and its digitisation as a realtively early historical newspaper archive. Only by becoming intimately familiar with the archive from which we draw our material can we develop and use tools that fit the data, instead of making the data fit the tool. Such source criticism is crucial to using these archives in a historically sound way.

In this same chapter, we began to address the question of computational resources as limiting factors in the design of research, which were further developed in the subsequent chapters. It found the influence of the computational power of the article retrieval step was significant in shaping the scope of the research questions, as it determined which keywords could be used. It forced the

balance between speculative queries in interesting directions on one hand and reliable keywords expected by historiography on the other to err on the side of caution. For these operations, speed was the deciding limiter. The computational limitations were further explored in chapters three and four in their influence on the use of specific tools; these were also found to be of great import, mainly by limiting the amount of data that could be processed in a single operation. Here, the limiting factor mainly related to the amount of memory required.

The question of whether it is possible for a single researcher to build their own tools, and if doing so provided any tangible benefit, was dealt with in chapters three and four. Here, this thesis conclusively proved that it is possible to produce tools tailored to a digital archive, by producing two functional tools of its own. One of these relied on a proven methodology, topic modelling through LDA, while the other was wholly new and untested. Building tools not only allows the historian using them to gain a better understanding of the way they work, but it also allows the development of entirely new approaches to the data. This thesis has shown through the development of these tools that one of the advantages to this approach is not being constrained by the tools that (professional) developers create; this allows researchers to ask their own questions, instead of being restricted to the ones their tools or archives allow them to ask. Two concrete cases of this for the project were the ability to construct a topic viewer that linked back to the original documents, and the construction of the visualisation tool *pro se*.

These two chapters also addressed specific questions concerning the methodology of using the two tools. Firstly, it found the epistemological grounding

of topic modelling, forcing the conclusion that while it is a powerful tool, because of its disassembly of the sources into first bags-of-words and subsequently probability distributions over the corpus of those bags, in a historical context the distance to the source becomes problematic. In addition to these issues, it was found that topic models suffer significantly from OCR transcription errors, forcing them to be generated using a relatively low number of topics. This inevitably means the conclusions that can be drawn from these models have to be general in nature. For these two reasons, this thesis advises topic models are treated with epistemological scepticism when used for historical research.

This thesis has shown that spatial visualisation of newspaper articles is possible. It finds there are no epistemological issues underlying the creation of such visualisations; on the contrary, there exists a small but substantial body of theoretical underpinning and practical case studies which would benefit from this technique. The project found that the recreation of just a little of the physicality that was lost when the digitisation took place adds significantly to the understanding of the research subject, while prompting the researcher to ask a variety of new questions. The process this thesis uses for visualising newspaper spatiality does not impact the dimensionality of data the way topic modelling does; instead, it produces its data from the original by recombination. This makes it more versatile than topic modelling, as it can be used to ask questions of the archive, a set of articles from the archive, or an individual article (which topic modelling cannot do). It also means that it lacks the epistemological issues that topic modelling struggles with, as instead of dissociating the contents of the sources to construct an analytical space, spatial visualisation collates information from the

sources into its analysis. Spatial article visualisation offers the opportunity to ask research questions that were previously impossible and opens the door for many new avenues of research.

In its final chapter, this thesis engaged with the third of these problems: how to use tools for a sound study of historical sources. By applying the tools it had developed to specific historical research questions, it was able to show that they can be used when certain factors and limitations are kept in mind. Amongst these is a requirement to work within the frameworks offered by historiography, theory, and the historical method; an awareness of the limitations imposed by the dataset used; and an appreciation that a tool will produce imperfect results which need a human interpreter to make sense of them. Its inclusion within the historiographical frameworks means that the use of tools always has to allow for the human researcher to have the final say in accepting or rejecting the findings of the tool, and to be transparent in those decisions. These decisions may be based on the limitations imposed by the data, and thus we need to closely engage with the data that we feed our tools, as these contain the only truths that the tool knows. Any errors or biases in the data will inevitably be reflected in the results of the tool. When these decisions are taken in the context of the data and the historiographical literature, weighted with proper scepticism, tools can make contributions to the historical debate.

The main contribution to the historiography on popular and banal imperialism is the support this thesis has provided for the argument that imperialism was widely present in the British press. It verified banal imperial

content amongst adverts, market news, personal notices, and in leisure newspapers, each of which played its own part in everyday flagging of the British empire. This thesis verified that these imperial elements were present from 1850 onwards, earlier than some other studies have theorised its presence. In the case studies, it used the topic modelling tool to identify several topics and concentrations of keywords which will be valuable avenues for future research.

The tools for topic modelling and spatial visualisation this thesis produced are able to generate genuine historical knowledge, so long as they are used within their limits. There will always be certain questions that these tools cannot answer. This project has found some of these limits for these tools on this dataset. Topic models in particular have been found to have some firm limits to their use. Their reliance on word collocation means that any error in article segmentation can cause errors in the topics that form. This problem is exacerbated by the OCR error rates in the archive, which also hide the actual word collocation frequencies with misspelled noise. These two factors result in a topic model that works best with relatively few topics, meaning it is unlikely to produce the extremely specific topics that the tool is capable of in ideal circumstances.

Despite these issues, topic models have been found to be useful tools for gesturing towards avenues of research by highlighting divergent or unexpected patterns in textual use of selected language. They provide the researcher with the 'hook' for further work, exposing entire categories of content that may have remained unnoticed in its banality. In addition to this exploratory use, topic models can, albeit cautiously, be used for measuring the relative size of a specific lexical

context – that is, the context in which a word is used. It can track this usage over time with reasonable accuracy, while providing a more substantial basis for the validity of its measurements than an approach of only keyword frequency measurement. As such, they are well suited for high-level and general research questions.

The large-scale visualisation of spatial relationships of newspaper articles is entirely new, and was also tested on the case studies. By being able to provide a density map of the spaces occupied by articles containing certain terms, it opens up entirely new possibilities for research. Because of this, at present the tool is surrounded by significant question marks: how to understand the spaces in Victorian newspapers through the eyes of a Victorian reader or editor; what were the conventions on article placement across the newspaper industry at the time; does the meaning of spaces change once illustrations become more common? All these theoretical questions are for future research. The development of a tool which allows for the visualisation of article placement does allow a new range of questions to be asked, and a new dimension of newspapers to be explored more systematically than was possible before. It allows for the article to be removed from the constraints of the single issue imposed by the archive and to re-emerge in the context of the patters of repetition that it originally existed within.

Research Implications

These findings on the usability of the concrete tools developed by this project all build towards a greater narrative surrounding the role that tools play in our research, and the ways that we as researchers interact with these tools. It emerged

in part out of a personal reaction against the environment of the Digital Humanities MA programme that I undertook, where an unbridled optimism about digital tools for humanities scholars seemed to abound. The general feeling there was that if we had more data, faster computers, and better algorithms, we could come to a more complete insight about the human condition. This thesis started as a project to critically examine the idea that such a future was possible or even desirable.

This project has taken it as given that all tools are limited, by design and by nature. By design, they are only capable of registering those things they are meant to, and only to operate within the limits they are designed to handle. As a thought experiment, consider applying a tool designed to organise a dataset, such as a topic modelling program, to a collection of truly random data. It will fail to create any meaningful order, as none existed within the data, but most tools would still give *some* answer, preserving an illusion of order. This is one key issue with digital tools: they often do not 'fail safe', explicitly avoiding answering a question if an answer cannot be found; instead they 'fail dumb' and give a tenuous answer that is virtually indistinguishable from a confident answer. Tools are also limited by the archive, as they can only study that which was recorded in the first place. Concrete examples for this project and archive include the ephemeral advertising wrappers and pages cut from the newspapers when they were bound in the nineteenth century, or the physical size of the page which the digital image represents. These lacunae in the data shape the tool and the research.

If we accept that these limits exist, it raises the question: how do we as humanities researchers deal with them, and gain the requisite knowledge to make

informed choices when we use them? I believe the best approach is by building our own tools. This does not mean constructing tools from the ground up: using packages and components that others have made is the only way to put together a tool within a reasonable timeframe; indeed, perhaps 'assembling' is a better term. This assembly should be guided by three factors: (1) a knowledge of the algorithm, (2) a knowledge of the data, and (3) a historical research question to guide the development process. We need to understand the algorithm that we wish to implement because each one has its own underlying assumptions and biases. For example, LDA assumes that each topic has its own distinct vocabulary – and uses that assumption to function. This is closely related to a knowledge of the data. If in a particular archive this assumption about vocabularies were to be false, it would require the researcher to choose a different algorithm or a different dataset. Finally, assembling a tool for the explicit purpose of answering a research question ties the knowledge of the algorithm and data together with practicality. It ensures the tool chosen is the right one to address the questions at hand.

My belief based on experiences with tools in chapter five is that tools can only ever be ancillary to other methods of epistemology, as they don't create knowledge, but can only facilitate discovery. A tool knows nothing but the information it has been given by its user, and has no way to verify that these ground truths are correct. All the information is already in the system, the tools simply reshape and reorder it, in order to allow the human researcher to discover which pieces can be used to gain knowledge. This is not to say that they should not be used, as they do have legitimate and valuable uses. However, they need to be used in a responsible manner.

Consequently, the Digital Humanities, as creators of tools, are best understood as an ancillary science to History. I realise this is a controversial statement about a discipline that is so broad and multi-facetted that it is still debating where its own boundaries lie, but I consider the model of the ancillary historical sciences and academic history appropriate, at least for the way this project has interacted with the field. As with other ancillary fields, such as the palaeographer transcribing the medieval text, the digital humanities supplies a skill needed to answer a specific historical research question. In the case of the digital humanities, this is the skill of providing a tool. Digital Humanities has its own specialised practices and academic infrastructures, which run in parallel to their historical counterparts. The key difference with other ancillary sciences is that DH also produces its own research and supplies its own research questions, however, this is not unheard of; chemistry acts as an ancillary science to art history if it analyses a pigment, but also produces research on its own. However, all of this does not absolve historians from the need to familiarise themselves on some basic level with the digital research environment they now inhabit. Even a historian not actively using tools should be able to understand the basic principles behind the sound deployment of tools in a research context. Historians need to become 'digitally tool-literate' to be able to engage with the work of the tool-builders and tool-users.

These beliefs are founded upon my experiences with this project. The way it sits between academic disciplines, and the resulting questions about the academic identity of the project have shaped it, but also forced it to ask questions about the way it uses techniques from these fields. They are based on the ways that tool, data,

and research question interacted and occasionally fought for dominance as the focus of the project, and upon the way the process of building the tools myself forced me to balance all three forces. This process has inspired a healthy epistemological scepticism towards tools and their findings. However, this is by no means a call to stop using tools or digital research. This project has shown how to use these tools responsibly and successfully.

In the end, our use of any tool has to be guided by a realisation that no matter how well-crafted, how powerful, or how advanced it is; if it is used outside of its tolerances, or if it is used to measure something it was not designed to, or if it is asked to measure data that it does not have, it will not produce valid results. This was true for any tool in history, and it will remain true into the future. "Instruments register only those things they're designed to register. Space still contains infinite unknowns."[1] If historians are aware of the limits of their tools then, as this project has also shown, they have the potential to significantly enrich our research and show us avenues of investigation. Like sculptors who are introduced to new techniques, styles, materials and tools, and well-acquainted with the intricacies of all of these, this will allow them the opportunity to carve new shapes in stone.

[1] Dr. Spock - M. Daniels & G. Rodenberry, 'The Naked Time', *Star Trek*, 1966.

Bibliography

Primary Sources

'Advertisements and Notices', *Manchester Mercury* (Manchester, 10 July 1880), p. 7

'Adverts and Notices', *Preston Chronicle* (Preston, 18 November 1882), p. 4

Anonymous, 'Facts for Electors to Remember', *Huddersfield Chronicle* (Huddersfield, 27 August 1885), p. 3

Coghlan, Sir Timothy Augustine, *A Statistical Account of the Seven Colonies of Australasia, 1895-6*, New South Wales Bureau of Statistics and Economics (Sydney, Australia: Potter, 1896)

'Death in a Snowdrift', *Pall Mall Gazette* (London, 18 September 1890), p. 6

'Educational Advertisements', *Hampshire Post and Telegraph* (Portsmouth, 1 May 1880), p. 2

'Family Notices', *Preston Chronicle* (Preston, 30 November 1850), p. 3

'Foreign Intelligence', *Reynold's Newspaper* (London, 17 January 1869), p. 8

[G.W.M. Reynolds], 'Notices to Correspondents', *Reynold's Newspaper* (London, 25 May 1851), p. 8

———, 'Notices to Correspondents', *Reynold's Newspaper* (London, 31 October 1852), p. 8

———, 'Notices to Correspondents', *Reynold's Newspaper* (London, 12 November 1854), p. 8

———, 'Notices to Correspondents', *Reynold's Newspaper* (London, 10 March 1852), p. 8

'Ladies Outfitting, Oldham & Sons.', *Freeman's Journal* (Dublin, 30 March 1877), p. 1

'Our Government of India, and the Famine', *Sheffield & Rotherham Independent* (Sheffield, 25 August 1877), p. 2

'Political and Social: Notes and Comments', *The Examiner* (London, 2 February 1878), pp. 10-11 (138-139)

'Small Dose, Small Pill, Small Price: Clarke's Pills', *Western Mail* (Cardiff, 21 August 1890)

'Southsea Visitors', *Hampshire Post and Telegraph* (Portsmouth, 7 August 1882), p. 3

'The Indo-European Telegraph.; Interesting Details of Laying the Line Difficulties with the Natives. the Manufacture of the Cable. Conveying the Cable. the First Station. Among the Arabs. Coming to Terms" with the Sheiks. Natural Difficulties. a Surprise. Dangers of Mud. the Last Obstacle.', *The New York Times*, 26 March 1865, <https://www.nytimes.com/1865/03/26/archives/the-indoeuropean-telegraph-interesting-details-of-laying-the-line.html> [accessed 19 November 2019]

'The Massacre of Missionaries', *Pall Mall Gazette* (London, 21 September 1894), p. 6

'The Perilous State of the Atlantic', *Pall Mall Gazette* (London, 21 September 1894), p. 8

'The Retention of Candahar', *Southampton Herald* (Southampton, 14 September 1880), p. 2

'The Southsea Visitors' List', *Hampshire Post and Telegraph* (Portsmouth, 30 May 1868), p. 6

'The Southsea Visitors' List', *Hampshire Post and Telegraph* (Portsmouth, 25 July 1868), p. 6

'Walker & Sons. - India, China & Ceylon Teas', *Aberdeen Journal* (Aberdeen, 22 July 1898), p. 4

Literature

'300 Years of the British Press Goes Digital – Gale and the British Library Build a Digital Reading Room', *Gale-Cengage Press Releases*, 2008 <https://news.cengage.com/library-research/300-years-of-the-british-press-goes-digital-%e2%80%93-gale-and-the-british-library-build-a-digital-reading-room/> [accessed 18 August 2019]

Acheson, Graeme G., and John D. Turner, 'The Secondary Market for Bank Shares in Nineteenth-Century Britain', *Financial History Review*, 15.2 (2008), 123–51 <https://doi.org/10.1017/S0968565008000139>

Aizawa, Akiko, 'An Information-Theoretic Perspective of Tf–Idf Measures', *Information Processing & Management*, 39.1 (2003), 45–65 <https://doi.org/10.1016/S0306-4573(02)00021-3>

Alex, Beatrice, and John Burns, 'Estimating and Rating the Quality of Optically Character Recognised Text', in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14 (New York, NY, USA: ACM, 2014), pp. 97–102 <https://doi.org/10.1145/2595188.2595214>

Allington, Daniel, Sarah Brouillette, and David Golumbia, 'Neoliberal Tools (and Archives): A Political History of Digital Humanities', *Los Angeles Review of Books* <https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/>

Alston, Rex, 'Sports: Games of Bat and Ball: Cricket', *Encylopedia Britannica* (Chicago & London: University of Chicago Press, 1985), 130–36

Anderson, Benedict, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, Rev. and extended ed. (London: Verso, 1991)

Arguing with Digital History Working Group, *Digital History and Argument* (Roy Rosenzweig Center for History and New Media, 13 November 2017) <https://rrchnm.org/wordpress/wp-content/uploads/2017/11/digital-history-and-argument.RRCHNM.pdf>

Asuncion, Arthur, Max Welling, Padhraic Smyth, and Yee Whye Teh, 'On Smoothing and Inference for Topic Models', *ArXiv:1205.2662 [Cs, Stat]*, 2012 <http://arxiv.org/abs/1205.2662>

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, 'Synthesizing Robust Adversarial Examples', *ArXiv:1707.07397 [Cs]*, 2018 <http://arxiv.org/abs/1707.07397>

Baker, William J., 'The Making of a Working-Class Football Culture in Victorian England', *Journal of Social History*, 13.2 (1979), 241–51

Barker, Hannah, and Simon Burrows, *Press, Politics and the Public Sphere in Europe and North America, 1760-1820* (Cambridge: Cambridge University Press, 2002)

Barnhurst, Kevin G., and John Nerone, *The Form of News: A History* (New York: Guilford Press, 2002)

———, 'Design Trends in US Front Pages, 1885–1985', *Journalism Quarterly*, 68.4 (1991), 796–804

Barton, Roger Neil, 'New Media: The Birth of Telegraphic News in Britain 1847–68', *Media History*, 16.4 (2010), 379–406 <https://doi.org/10.1080/13688804.2010.507475>

Bassiou, Nikoletta K., and Constantine L. Kotropoulos, 'Online PLSA: Batch Updating Techniques Including Out-of-Vocabulary Words', *IEEE Transactions on Neural Networks and Learning Systems*, 25.11 (2014), 1953–66 <https://doi.org/10.1109/TNNLS.2014.2299806>

Beals, M., 'Scissors and Paste: The Georgian Reprints, 1800–1837', *Journal of Open Humanities Data*, 3.0 (2017), 1 <https://doi.org/10.5334/johd.8>

———, 'Anatomy of a Newspaper: The Caledonian Mercury, 20 June 1825', *MHBeals.Com*, 2017 <http://mhbeals.com/anatomy-of-a-newspaper-the-caledonian-mercury-20-june-1825/> [accessed 22 October 2019]

———, 'Close Readings of Big Data: Triangulating Patterns of Textual Reappearance and Attribution in the Caledonian Mercury, 1820-1840', *Victorian Periodicals Review*, 51.4 (2018), 616–39

Beals, M. H., and Emily Bell, 'British Library 19th Century Newspapers', in *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges* (Loughborough, 2020) <10.6084/m9.figshare.11560059.>

Bellegarda, Jerome R, 'Statistical Language Model Adaptation: Review and Perspectives', *Speech Communication*, 42.1 (2004), 93–108 <https://doi.org/10.1016/j.specom.2003.08.002>

Benveniste, Emile, 'Subjectivity in Language', *Problems in General Linguistics*, 1 (1971), 223–230

Benwell, Bethan, and Elisabeth Stokoe, 'Theorising Discourse and Identity', in *Discourse and Identity* (Edinburgh: Edinburgh University Press, 2006)

van den Berg, Hein, Arianna Betti, Thom Castermans, Rob Koopman, Bettina Speckmann, Kevin Verbeek, and others, 'A Philosophical Perspective on Visualization for Digital Humanities', in *3rd Workshop on Visualization for the Digital Humanities* (presented at the IEEE VIS 2018, Berlin, 2018) <http://vis4dh.dbvis.de/papers/2018/A%20Philosophical%20Perspective%20on%20Visualization%20for%20Digital%20Humanities.pdf>

Berger, Stefan, Kevin Passmore, and Heiko Feldner, *Writing History: Theory & Practice*, Writing History, 2nd ed. (London: Bloomsbury Academic, 2010)

Berridge, Virginia. S., 'Popular Journalism and Working Class Attitudes 1854-1886 : A Study of Reynold's Newspaper, Lloyd's Weekly Newspaper and the Weekly Times.' (unpublished Ph.D., Birkbeck (University of London), 1976)

Berry, David M., 'The Computational Turn: Thinking About the Digital Humanities', *Culture Machine*, 12 (2011), 23

Berry, Michael W., Susan T. Dumais, and Amy T. Shippy, *A Case Study of Latent Semantic Indexing* (Tennessee: University of Tennessee, 1995)

Billig, Michael, *Banal Nationalism* (London: SAGE, 1995)

Bingham, Adrian, '"The Digitization of Newspaper Archives: Opportunities and Challenges for Historians"', *Twentieth Century British History*, 21.2 (2010), 225–31 <https://doi.org/10.1093/tcbh/hwq007>

Bird, Steven, Ewan Klein, and Edward Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (Sebastopol, CA: O'Reilly Media, Inc., 2009)

Birdell, T.A., 'The British Museum Duplicate Sales, 1769-1832, and Their Significance for the Early Collections', in *Libraries Within the Library: The Origins of the British Library's Printed Collections*, ed. by Giles Mandelbrote and Barry Taylor (London: British Library, 2009)

Birns, Nicholas, 'Daily Telegraph', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 158–59

Black, Jeremy, review of *Visions of Empire: Patriotism, Popular Culture and the City, 1870–1939; Britain's Imperial Muse: The Classics, Imperialism, and the Indian Empire, 1784–1914; Crisis in the Mediterranean: Naval Competition and Great Power Politics, 1904–1914; Mapping the End of Empire: American and British Strategic Visions in the Postwar World; The Second British Empire: In the Crucible of the Twentieth Century; Distant Strangers: How Britain Became Modern*, by Brad Beaven, Christopher Hagerman, Jon K. Hendrickson, Aiyaz Husain, Timothy H. Parsons, and James Vernon, *Journal of World History*, 26.2 (2016), 395–400 <https://doi.org/10.1353/jwh.2016.0041>

Blei, David, *Prof. David Blei - Probabilistic Topic Models and User Behavior* (Edinburgh, 2017) <https://www.youtube.com/watch?v=FkckgwMHP2s&t=1086s> [accessed 24 February 2020]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3.Jan (2003), 993–1022

Bleske, Glen L., 'Ms. Gates Takes over: An Updated Version of a 1949 Case Study', *Newspaper Research Journal*, 12.4 (1991), 88–97 <https://doi.org/10.1177/073953299101200409>

Borland, David, and Russell M. Taylor, 'Rainbow Color Map (Still) Considered Harmful', *IEEE Computer Graphics and Applications*, 27.2 (2007), 14–17 <https://doi.org/10.1109/MCG.2007.323435>

Bos, Maarten van den, and Hermione Giffard, 'Mining Public Discourse for Emerging Dutch Nationalism', *Digital Humanities Quarterly*, 10.3 (2016)

Bratton, J. S., 'Of England, Home, an Duty: The Image of England in Victorian and Edwardian Juvinile Fiction', in *Imperialism and Popular Culture*, by John M. MacKenzie (Manchester: Manchester University Press, 1986), pp. 73–93

Brauer, René, Mirek Dymitrow, and Mats Fridlund, 'The Digital Shaping of Humanities Research: The Emergence of Topic Modeling within Historical Studies', in *Enacting Futures: DASTS 2014*, 2014

Breisach, Ernst, *Historiography: Ancient, Medieval, and Modern*, 3rd edn (Chicago & London: University of Chicago Press, 2007)

'British Library 19th Century Newspapers : JISC', 2008 <https://web.archive.org/web/20080918080147/http://www.jisc.ac.uk/whatwedo/programmes/digitisation/bln> [accessed 18 August 2019]

'British Library 19th-Century Newspapers : JISC', 2010 <https://web.archive.org/web/20100607033152/http://www.jisc.ac.uk:80/whatwedo/programmes/digitisation/bln> [accessed 18 August 2019]

'British Library and Brightsolid Partnership to Digitise up to 40 Million Pages of Historic Newspapers', *British Newspaper Archive Press and Media Information*, 2011 <https://web.archive.org/web/20110526213116/http://www.britishnewspaperarchive.co.uk/archive-media.php> [accessed 18 August 2019]

British Library Newspapers (Gale Cengage Learning) <https://www.gale.com/binaries/content/assets/gale-us-en/primary-sources/newsvault/gps_newsvault_britishlibrarynewspapers_academic_brochure_all.pdf>

Broadwell, Peter, Tomoko Bialock, and Hiroyuki Ikuurada, 'Macroscopic Exploration of Large Text and Image Collections via Similarity Heatmaps', in *Proceedings of the 7th Conference of Japanese Association for Digital Humanities* (presented at the JADH 2017, Doshisha, 2017), pp. 1–4

Brodie, Allan, 'The Brown Family Adventure – Seaside Holidays in Kent in the Mid-Nineteenth Century', *Journal of Tourism History*, 5.1 (2013), 1–24 <https://doi.org/10.1080/1755182X.2012.758671>

Brody, David Eric, 'Building Empire: Architecture and American Imperialism in the Philippines', *Journal of Asian American Studies*, 4.2 (2001), 123–45 <https://doi.org/10.1353/jaas.2001.0013>

Bromley, Michael, and Tom O'Malley, 'Introduction', in *A Journalism Reader* (Psychology Press, 1997)

Buettner, Elisabeth, *Empire Families: Britons and Late Imperial India* (Oxford: Oxford University Press, 2004)

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp, *Digital_Humanities* (Cambridge, MA: MIT Press, 2016)

Burke, Peter, *What Is Cultural History?*, 2nd edn (New Jersey: Wiley, 2008)

Burniske, R. W., *Literacy in the Digital Age* (Thousand Oaks, CA: Corwin Press, 2008)

Busa, R., 'The Annals of Humanities Computing: The Index Thomisticus', *Computers and the Humanities*, 14.2 (1980), 83–90 <https://doi.org/10.1007/BF02403798>

Buzzetti, Dino, 'The Origins of Humanities Computing and the Digital Humanities Turn', *Humanist Studies & the Digital Age*, 6.1 (2019), 32–58

Cannadine, David, *Ornamentalism: How the British Saw Their Empire* (London: Penguin Books, 2001)

———, *Victorious Century: The United Kingdom, 1800-1906* (London: Penguin, 2018)

Champion, Erik Malcolm, 'Digital Humanities Is Text Heavy, Visualization Light, and Simulation Poor', *Digital Scholarship in the Humanities*, 32.suppl_1 (2017), i25–32 <https://doi.org/10.1093/llc/fqw053>

Chaney, Allison June-Barlow, and David M. Blei, 'Visualizing Topic Models', in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012 <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4645>

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei, 'Reading Tea Leaves: How Humans Interpret Topic Models', in *Advances in Neural Information Processing Systems*, 2009, pp. 288–296

Chen, Berlin, 'Latent Topic Modelling of Word Co-Occurence Information for Spoken Document Retrieval', in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3961–64 <https://doi.org/10.1109/ICASSP.2009.4960495>

Cheradame, H., S. Ipert, and E. Rousset, 'Mass Deacidification of Paper and Books. I: Study of the Limitations of the Gas Phase Processes', *Restaurator*, 24.4 (2008), 227–239 <https://doi.org/10.1515/REST.2003.227>

Chez, Kerry, 'Daily Mail', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 157–58

Chilton, Lisa, *Agents of Empire: British Female Migration to Canada and Australia, 1860-1930* (Toronto: University of Toronto Press, 2007)

Chivers, Ian, and Jane Sleightholme, 'An Introduction to Algorithms and the Big O Notation', in *Introduction to Programming with Fortran: With Coverage of Fortran 90, 95, 2003, 2008 and 77*, ed. by Ian Chivers and Jane Sleightholme (Springer International Publishing, 2015), pp. 359–64

Choi, Hyunyoung, and Hal R. Varian, *Predicting the Present with Google Trends* (Rochester, NY: Social Science Research Network, 1 June 2012) <https://doi.org/10.1111/j.1475-4932.2012.00809.x>

Chuang, Jason, Christopher D. Manning, and Jeffrey Heer, 'Termite: Visualization Techniques for Assessing Textual Topic Models', in *Proceedings of the International Working Conference on Advanced Visual Interfaces* (ACM, 2012), pp. 74–77

Church, Roy, 'Advertising Consumer Goods in Nineteenth-Century Britain: Reinterpretations', *The Economic History Review*, 53.4 (2000), 621–45

Cioffi-Revilla, Claudio, *Introduction to Computational Social Science: Principles and Applications*, Texts in Computer Science, 2nd edn (Springer International Publishing, 2017)

Cleall, Esme, Laura Ishiguro, and Emily J. Manktelow, 'Imperial Relations: Histories of Family in the British Empire', *Journal of Colonialism and Colonial History*, 14.1 (2013)

Colley, Linda, *Britons: Forging the Nation 1707-1837*, 2nd edn (New Haven and London: Yale University Press, 2014)

Colville, Quintin, 'Enacted and Re-Enacted in Life and Letters: The Identity of the Jack Tar, 1930 to Date', *Journal for Maritime Research*, 18.1 (2016), 37–53 <https://doi.org/10.1080/21533369.2016.1172840>

Conley, Mary, *From Jack Tar to Union Jack: Representing Naval Manhood in the British Empire, 1870-1918* (Manchester: Manchester University Press, 2009)

Couper, Mick, Frauke Kreuter, and Lars Lyberg, 'The Use of Paradata to Monitor and Manage Survey Data Collection', in *JSM Proceedings - Survey Research Methods Section* (Alexandria, VA: American Statistical Association, 2010), pp. 281–96 <http://www.asasrms.org/Proceedings/y2010/Files/306107_55863.pdf>

Crowe, Charles, 'Time on the Cross: The Historical Monograph as a Pop Event', *The History Teacher*, 9.4 (1976), 588–630 <https://doi.org/10.2307/492099>

Da, Nan Z., 'The Computational Case against Computational Literary Studies', *Critical Inquiry*, 45.3 (2019), 601–39 <https://doi.org/10.1086/702594>

Daniels, M., and G. Roddenberry, 'The Naked Time', *Star Trek*, 1966

Danielson, Wayne A., and James J. Mullen, 'A Basic Space Unit for Newspaper Content Analysis', *Journalism Quarterly*, 42.1 (1965), 108–10 <https://doi.org/10.1177/107769906504200114>

Darian-Smith, Kate, Patricia Grimshaw, and Stewart Macintyre, eds., *Britishness Abroad: Transnational Movements and Imperial Cultures* (Melbourne: Melbourne University Press, 2007)

Darwin, John, *Unfinished Empire: The Global Expansion of Britain* (London: Penguin Books, 2013)

Daume, Stefan, Matthias Albert, and Klaus von Gadow, 'Assessing Citizen Science Opportunities in Forest Monitoring Using Probabilistic Topic Modelling', *Forest Ecosystems*, 1.1 (2014), 11 <https://doi.org/10.1186/s40663-014-0011-6>

David, Paul A., Herbert G. Gutman, Richard Sutch, Peter Temin, and Gavin Wright, *Reconing with Slavery* (Oxford University Press, 1985)

Dawson, Graham, *Soldier Heroes: British Adventure, Empire, and the Imagining of Masculinities* (London: Routledge, 1994)

Day, John C., 'The Library Scene in an English City: Newcastle-on-Tyne Libraries, 1850-2000', in *Cambridge History of Libraries in Britain and Ireland* (Cambridge: Cambridge University Press, 2006), III, 206–15

De Fina, Anna, and Sabina Perrino, '"Transnational Identities"', *Applied Linguistics*, 34.5 (2013), 509–15 <https://doi.org/10.1093/applin/amt024>

De Schryver, Reginald, *Historiografie: Vijfentwintig Eeuwen Geschiedschrijving van West-Europa*, Ancorae, 8, 3rd edn (Leuven: Universitaire Pers Leuven, 1997)

De Waal, A., and E. Barnard, 'Evaluating Topic Models with Stability', in *Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa* (presented at the PRASA 2008, Cape Town, South Africa, 2008), pp. 79–84

Deerwester, Scott C., Susan T. Dumais, George W. Furnas, Richard A. Harshman, Thomas K. Landauer, Karen E. Lochbaum, and others, 'Computer Information Retrieval Using Latent Semantic Structure', 1989

———, 'Indexing by Latent Semantic Analysis', *Journal of the American Society for Information Science*, 41.6 (1990), 391–407

Demoor, Marysa, and Laurel Brake, eds., *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009)

DiMaggio, Paul, Manish Nag, and David Blei, 'Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding', *Poetics*, 41.6 (2013), 570–606 <https://doi.org/10.1016/j.poetic.2013.08.004>

Dray, W. H., 'On the Nature and Role of Narrative in Historiography', *History and Theory*, 10.2 (1971), 153–71 <https://doi.org/10.2307/2504290>

Dumais, Susan T., 'Latent Semantic Analysis', *Annual Review of Information Science and Technology*, 38.1 (2004), 188–230 <https://doi.org/10.1002/aris.1440380105>

Dupuy, R.E., and T.N. Dupuy, *The Collins Encyclopedia of Military History: From 3500 BC to the Present*, 4th edn (New York: HarperCollins Publishers Ltd, 1993)

Dutt, Romesh C., *The Economic History of India in the Victorian Age: From the Accession of Queen Victoria in 1837 to the Commencement of the Twentieth Century*, II vols (s.l.: K. Paul, Trench, Trübner & Company Limited, 1904)

Edmond, Jennifer, 'Managing Uncertainty in the Humanities: Digital and Analogue Approaches', in *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18 (Salamanca, Spain: Association for Computing Machinery, 2018), pp. 840–844 <https://doi.org/10.1145/3284179.3284326>

Eijnatten, Joris van, Toine Pieters, and Jaap Verheul, 'Big Data for Global History: The Transformative Promise of Digital Humanities', *BMGN - Low Countries Historical Review*, 128.4 (2013), 55–77 <https://doi.org/10.18352/bmgn-lchr.9350>

Eldridge II, Scott, 'Change and Continuity: Historicizing the Emergence of Online Media', in *The Routledge Companion to British Media History*, ed. by Martin Conboy and John Steel (Routledge, 2014), pp. 528–38

Farish, Matthew, 'Modern Witnesses: Foreign Correspondents, Geopolitical Vision, and the First World War', *Transactions of the Institute of British Geographers*, 26.3 (2001), 273–87 <https://doi.org/10.1111/1475-5661.00022>

Feely, Catherine, '"What Say You to Free Trade in Literature?" The Thief and the Politics of Piracy in the 1830s', *Journal of Victorian Culture*, 19.4 (2014), 497–506 <https://doi.org/10.1080/13555502.2014.967545>

Ferguson, Niall, *Empire: How Britain Made the Modern World* (London: Penguin, 2008)

Fetahu, Besnik, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl, 'A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles', in *The Semantic Web: Trends and Challenges*, ed. by Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2014), pp. 519–34 <https://doi.org/10.1007/978-3-319-07443-6_35>

Fieldhouse, D. K., 'Gentlemen, Capitalists, and the British Empire', *The Journal of Imperial and Commonwealth History*, 22.3 (1994), 531–41 <https://doi.org/10.1080/03086539408582938>

Finlayson, Samuel G., Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam, 'Adversarial Attacks Against Medical Deep Learning Systems', *ArXiv:1804.05296 [Cs, Stat]*, 2019 <http://arxiv.org/abs/1804.05296> [accessed 25 February 2020]

Fleming, N. C., review of *Britain's Experience of Empire in the Twentieth Century*, by Andrew Thomson, *The Journal of Imperial and Commonwealth History*, 41.3 (2013), 536–37 <https://doi.org/10.1080/03086534.2013.823745>

Fletcher, Charles R., and Brian Linzie, 'Motive and Opportunity: Some Comments on LSA, HAL, KDC, and Principal Components', *Discourse Processes*, 25.2–3 (1998), 355–61 <https://doi.org/10.1080/01638539809545032>

Floud, Roderick, and Paul Johnson, eds., *The Cambridge Economic History of Britain*, 2 vols (Cambridge: Cambridge University Press, 2014)

Fogel, Robert W., and Stanley L. Engerman, *Time on the Cross: The Economics of American Negro Slavery* (London and New York: W. W. Norton, 1995)

Freeman, Mark, Robin Pearson, and James Taylor, '"A Doe in the City": Women Shareholders in Eighteenth- and Early Nineteenth-Century Britain', *Accounting, Business & Financial History*, 16.2 (2006), 265–91 <https://doi.org/10.1080/09585200600756282>

Friendly, Michael, 'A Brief History of Data Visualization', in *Handbook of Data Visualization*, by Chun-houh Chen, Wolfgang Härdle, and Antony Unwin (Berlin, Heidelberg: Springer Berlin Heidelberg, 2008), pp. 15–56 <https://doi.org/10.1007/978-3-540-33037-0_2>

———, 'Milestones in the History of Data Visualization: A Case Study in Statistical Historiography', in *Classification — the Ubiquitous Challenge*, ed. by Claus Weihs and Wolfgang Gaul (Berlin/Heidelberg: Springer-Verlag, 2005), pp. 34–52 <https://doi.org/10.1007/3-540-28084-7_4>

———, 'Visions and Re-Visions of Charles Joseph Minard', *Journal of Educational and Behavioral Statistics*, 27.1 (2002), 31–51 <https://doi.org/10.3102/10769986027001031>

Frith, David, *England versus Australia: an Illustrated History of Every Test Match since 1877*, 12th edn (London: Viking, 2007)

———, *The Golden Age of Cricket, 1890-1914* (Guildford: Lutterworth Press : Richard Smart Pub., 1978)

Frost, W., 'Heritage, Nationalism, Identity: The 1861-62 England Cricket Tour of Australia', *The International Journal of the History of Sport*, 19.4 (2002), 55–69 <https://doi.org/10.1080/714001786>

Fyfe, Paul, 'Access, Computational Analysis, and Fair Use in the Digitized Nineteenth- Century Press', *Victorian Periodicals Review*, 51.4 (2018), 716–37

Gale-Cengage, 'New Product Enhancements Announced for Digital Scholar Lab', *Product Support*, 2019 </updates/dsl-feb19> [accessed 3 March 2020]

van Galen, Quintus, and Bob Nicholson, 'In Search of America: An Introduction to Topic Modelling Nineteenth-Century Newspaper Archives', *Digital Journalism*, 2018

van Gansen, Kristof, '"Une Page Est Une Image.": Tekst Als Beeld in Arts et Métiers Graphiques', *Tijdschrift Voor Tijdschriftstudies*, 35, 2014, 5–21

García, Mario R., *Contemporary Newspaper Design: A Structural Approach* (Prentice-Hall, 1987)

Gates, David, *Warfare in the Nineteenth Century*, European History in Perspective, 7 (Basingstoke: Palgrave Macmillan, 2001)

Gibbs, Fred, and Trevor Owens, 'Building Better Digital Humanities Tools', *DH Quarterly*, 6.2 (2012)

Glassock, Simon, 'Good Sports? Scotland, Empire and Rugby c.1924–1928', *Sport in History*, 36.3 (2016), 350–69

Goldstone, Andrew, *Dfr-Browser*, version 0.8.1, 2019 <https://github.com/agoldst/dfr-browser> [accessed 4 March 2020]

Golub, G. H., and C. Reinsch, 'Singular Value Decomposition and Least Squares Solutions', in *Linear Algebra*, ed. by J. H. Wilkinson, C. Reinsch, and F. L. Bauer, Handbook for Automatic Computation (Berlin, Heidelberg: Springer, 1971), pp. 134–51 <https://doi.org/10.1007/978-3-662-39778-7_10>

Goodman, Jordan, *Tobacco in History: The Cultures of Dependence* (London and New York: Routledge, 2005)

Gould, Peter, 'Portsmouth Corporation Transport: 1901-1986', *Portsmouth Corporation Transport, 1901-1986*, 1999 <http://www.petergould.co.uk/local_transport_history/fleetlists/portsmouth1.htm> [accessed 18 June 2019]

Graham, Shawn, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (World Scientific Publishing Company, 2015)

Grusin, Richard, 'The Dark Side of Digital Humanities: Dispatches from Two Recent Mla Conventions', *Differences*, 25.1 (2014), 79–92 <https://doi.org/10.1215/10407391-2420009>

Guiliano, Jennifer, 'Toward a Praxis of Critical Digital Sport History', *Journal of Sport History*, 44.2 (2017), 146–59 <https://doi.org/10.5406/jsporthistory.44.2.0146>

Gutman, Herbert George, *Slavery and the Numbers Game: A Critique of Time on the Cross* (University of Illinois Press, 1975)

Guttmann, A., *Games and Empires: Modern Sports and Cultural Imperialism* (New York: Columbia University Press, 1994)

Gwinn, Nancy E., 'The Fragility of Paper: Can Our Historical Record Be Saved?', *The Public Historian*, 13.3 (1991), 33–53 <https://doi.org/10.2307/3378551>

Hall, David, Daniel Jurafsky, and Christopher D. Manning, 'Studying the History of Ideas Using Topic Models', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2008), pp. 363–371 <http://dl.acm.org/citation.cfm?id=1613763>

Hall, Mark, 'Opportunities and Risks in Digital Humanities Research', in preprint

Harari, Yuval Noah, *Homo Deus: A Brief History of Tomorrow* (New York: HarperCollins Publishers Ltd, 2016)

Harris, P. R., *A History of the British Museum Library, 1753-1973* (London: British Library, 1998)

Harris, P.R., 'The British Museum Library, 1857-1973', in *Cambridge History of Libraries in Britain and Ireland* (Cambridge: Cambridge University Press, 2006), III, 281–98

Harvie, Christopher, 'Revolution and the Rule of Law (1789-1851)', in *The Oxford Illustrated History of Britain*, ed. by Kenneth O. Morgan, 2nd edn (Oxford: Oxford University Press, 2009), pp. 419–62

Haskell, Francis, *History and Its Images: Art and the Interpretation of the Past* (New Haven, CT: Yale University Press, 1993)

Hatton, Timothy J., 'Population, Migration, and Labour Supply: Great Britian 1871-2011', in *The Cambridge Economic History of Britain*, ed. by Roderick Floud and Paul Johnson, 2 vols (Cambridge: Cambridge University Press, 2014), II, 95–121

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier, 'An Unsupervised Neural Attention Model for Aspect Extraction', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (presented at the Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada: Association for

Computational Linguistics, 2017), pp. 388–97 <https://doi.org/10.18653/v1/P17-1036>

Healey, C.G., 'Choosing Effective Colours for Data Visualization', in *Proceedings of Seventh Annual IEEE Visualization '96*, 1996, pp. 263–70 <https://doi.org/10.1109/VISUAL.1996.568118>

Hedley, Alison, 'Advertisements, Hyper-Reading, and Fin de Siècle Consumer Culture in the Illustrated London News and the Graphic', *Victorian Periodicals Review*, 51.1 (2018), 138–67

Hendley, Matthew C., *Organized Patriotism and the Crucible of War: Popular Imperialism in Britain, 1914-1932* (Montreal: McGill-Queen's University Press, 2012)

Hess, Kristy, and Sarah Pinto, 'Forever In Our Hearts', *Media History*, 26.2 (2020), 105–21 <https://doi.org/10.1080/13688804.2018.1482205>

Hewitt, Martin, *The Dawn of the Cheap Press in Victorian Britain: The End of the 'Taxes on Knowledge', 1849-1869* (London and New York: Bloomsbury, 2014)

Hills, Thomas T., Eugenio Proto, Daniel Sgroi, and Chanuki Illushka Seresinhe, 'Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books', *Nature Human Behaviour*, 3.12 (2019), 1271–75 <https://doi.org/10.1038/s41562-019-0750-z>

'History of the Atlantic Cable & Submarine Telegraphy - Cable Timeline' <http://atlantic-cable.com//Cables/CableTimeLine/index1850.htm> [accessed 19 November 2019]

Hobbs, Andrew, *A Fleet Street in Every Town: The Provincial Press in England, 1855-1900* (Cambridge: Open Book Publishers, 2018)

———, 'The Reading World of a Provincial Town: Preston, Lancashire 1855-1900', in *The History of Reading, Volume 2: Evidence from the British Isles, c.1750-1950*, ed. by K. Halsey and W. Owens (Basingstoke: Palgrave Macmillan, 2011), pp. 121–38

Hobsbawm, Eric, *The Invention of Tradition* (Cambridge: Cambridge University Press, 1983)

Hoekstra, Rik, 'Data Scopes for Digital History Research', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2, 2018, 1–16 <https://doi.org/10.1080/01615440.2018.1484676>

Hofmann, Thomas, 'Probabilistic Latent Semantic Indexing', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99 (Berkeley, California, USA: Association for Computing Machinery, 1999), pp. 50–57 <https://doi.org/10.1145/312624.312649>

Holley, Rose, 'How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs', *D-Lib Magazine*, 15.3/4 (2009) <https://doi.org/10.1045/march2009-holley>

Hopson, John, 'The British Library and Its Antecedents', in *Cambridge History of Libraries in Britain and Ireland* (Cambridge: Cambridge University Press, 2006), III, 299–316

Houston, Natalie M., "Towards a Visual Analysis of Victorian Poetics', *Victorian Studies* 56.3 (2014), 498-510Howell, Philip, David Beckingham, and Francesca Moore, 'Managed Zones for Sex Workers in Liverpool: Contemporary Proposals, Victorian Parallels', *Transactions of the Institute of British Geographers*, 33.2 (2008), 233–50 <https://doi.org/10.1111/j.1475-5661.2008.00292.x>

Huggins, Mike, 'Culture, Class and Respectability: Racing and the English Middle Classes in the Nineteenth Century', *The International Journal of the History of Sport*, 11.1 (1994), 19–41 <https://doi.org/10.1080/09523369408713845>

Hutt, Allen, *The Changing Newspaper: Typographic Trends in Britain and America 1622-1972* (Gordon Fraser, 1973)

Itzkowitz, David C., 'Fair Enterprise or Extravagant Speculation: Investment, Speculation, and Gambling in Victorian England', *Victorian Studies*, 45.1 (2002), 121–47

Iwata, Tomoharu, Takeshi Yamada, and Naonori Ueda, 'Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents', in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08 (Las Vegas: Association for Computing Machinery, 2008), pp. 363–371 <https://doi.org/10.1145/1401890.1401937>

Jackson, Andrew J. H., 'Provincial Newspapers and the Development of Local Communities: The Creation of a Seaside Resort Newspaper for Ilfracombe, Devon, 1860–1', *Family & Community History*, 13.2 (2010), 101–13 <https://doi.org/10.1179/146311810X12851639314110>

Jackson, Kate, *George Newnes and the New Journalism in Britain, 1880–1910: Culture and Profit* (Burlington: Ashgate, 2001)

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers, 'Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling', *Digital Journalism*, 4.1 (2016), 89–106 <https://doi.org/10.1080/21670811.2015.1093271>

Jacobs, Thomas, and Robin Tschötschel, 'Topic Models Meet Discourse Analysis: A Quantitative Tool for a Qualitative Approach', *International Journal of Social*

Research Methodology, 22.5 (2019), 469–85 <https://doi.org/10.1080/13645579.2019.1576317>

James, Lawrence, *Raj: The Making of British India*, 2nd edn (London: Little, Brown and Company, 1998)

———, *The Rise and Fall of the British Empire*, 2nd edn (Abbacus: London, 1998)

Jänicke, Stefan, 'Valuable Research for Visualization and Digital Humanities: A Balancing Act', in *1st Workshop on Visualization for the Digital Humanities* (presented at the IEEE VIS 2016, Baltimore, USA, 2016) <http://vis4dh.dbvis.de/papers/2016/Valuable%20Research%20for%20Visualization%20and%20Digital%20Humanities%20A%20Balancing%20Act.pdf>

Jessop, M., 'Digital Visualization as a Scholarly Activity', *Literary and Linguistic Computing*, 23.3 (2008), 281–93 <https://doi.org/10.1093/llc/fqn016>

Jeurgens, Charles, 'The Scent of the Digital Archive: Dilemmas with Archive Digitisation', *BMGN-Low Countries Historical Review*, 128, 2013, 30–54

Kaufmann, Eric, 'Complexity and Nationalism', *Nations and Nationalism*, 23.1 (2017), 6–25 <https://doi.org/10.1111/nana.12270>

Kaul, Chandrika, *Reporting the Raj: The British Press and India, C. 1880-1922* (Manchester University Press, 2003)

Keim, Daniel A., and Daniela Oelke, 'Literature Fingerprinting: A New Method for Visual Literary Analysis', in *2007 IEEE Symposium on Visual Analytics Science and Technology* (presented at the 2007 IEEE Symposium on Visual Analytics Science and Technology, Sacramento, CA, USA: IEEE, 2007), pp. 115–22 <https://doi.org/10.1109/VAST.2007.4389004>

Kennedy, Dane, *Britain and Empire, 1880-1945* (London: Longman, 2002)

Kettunen, Kimmo, and Tuula Pääkkönen, 'Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means', 2016

King, Andrew, 'Advertising', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 5–7

King, Edmund, 'Digitisation of British Library Newspapers 1800-1900', *British Library Newspapers*, 2007 <http://find.galegroup.com/bncn/bncn_01.htm> [accessed 18 August 2019]

———, 'Digitisation of Newspapers at the British Library', *The Serials Librarian*, 49.1–2 (2005), 165–81 <https://doi.org/10.1300/J123v49n01_07>

Klijn, Edwin, 'The Current State-of-Art in Newspaper Digitization: A Market Perspective', *D-Lib Magazine*, 14.1/2 (2008) <https://doi.org/10.1045/january2008-klijn>

Knaplund, Paul, *The British Empire, 1815-1939* (London: Hamish Hamilton, 1942)

Knight, Robert, *The Indian Empire, and Our Financial Relations Therewith: A Papaer Read Before the London Indian Society* (London: Trubner & co., 1866)

Knuth, Donald E., 'Big Omicron and Big Omega and Big Theta', *ACM SIGACT News*, 8.2 (1976), 18–24 <https://doi.org/10.1145/1008328.1008329>

Koehn, Nancy Fowler, *The Power of Commerce: Economy and Governance in the First British Empire* (Cornell University Press, 1994)

Koistinen, Mika, Kimmo Kettunen, and Tuula Pääkkönen, 'Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing', in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 2017, pp. 277–283

Kolmer, Christian, 'Methods of Journalism Research - Content Analysis', in *Global Journalism Research: Theories, Methods, Findings, Future*, ed. by Martin Lèoffelholz and David Weaver (Oxford: Blackwell, 2007), pp. 117–30

Kong, Jing, Alex Scott, and Georg M. Goerg, 'Improving Topic Clustering on Search Queries with Word Co-Occurrence and Bipartite Graph Co-Clustering', 2016

Kontostathis, April, and William M. Pottenger, 'A Framework for Understanding Latent Semantic Indexing (LSI) Performance', *Information Processing & Management*, Formal Methods for Information Retrieval, 42.1 (2006), 56–73 <https://doi.org/10.1016/j.ipm.2004.11.007>

Koolen, Marijn, Jasmijn van Gorp, and Jacco van Ossenbruggen, 'Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice', *Digital Scholarship in the Humanities*, 34.2 (2019), 368–85 <https://doi.org/10.1093/llc/fqy048>

Kurvits, Roosmarii, 'The Visual Form of Estonian Newspapers from 1806 to 1940 and the Appearance Spiral Model', *Nordicom Review*, 29.2 (2008), 335–52 <https://doi.org/10.1515/nor-2017-0195>

———, 'The Visual Form of Newspapers as a Guide for Information Consumption', in *Things in Culture, Culture in Things*, ed. by A. Kannike and P. Laviolette, Approaches to Culture Theory, 3 (Tartu: University of Tartu Press, 2013), pp. 172–203

Киселев [Kiselev], А., 'История Оформления Русской Газеты (1702-1917 Гг.) [History of the Form of Russian Newspaper (1702-1917)]', in *The Visual*

Form of Estonian Newspapers from 1806 to 1940 and The Appearance Spiral Model, by Roosmarii Kurvits, Nordicom Review, 29, 2008, pp. 335–52

Lambie, James, *The Story of Your Life: A History of the Sporting Life Newspaper (1859-1998)* (Troubador Publishing Ltd, 2010)

Land, Isaac, *War, Nationalism and the British Sailor, 1750-1850* (Basingstoke: Palgrave Macmillan, 2009)

Land of Hope and Glory (London: Boosey & co., 1902)

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham, 'An Introduction to Latent Semantic Analysis', *Discourse Processes*, 25.2–3 (1998), 259–84 <https://doi.org/10.1080/01638539809545028>

Landauer, Thomas K., and Michael L. Littman, 'Computerized Cross-Language Document Retrieval Using Latent Semantic Indexing', 1994

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini, 'Content Analysis of 150 Years of British Periodicals', *Proceedings of the National Academy of Sciences*, 114.4 (2017), E457–65 <https://doi.org/10.1073/pnas.1606380114>

Lapacherie, J.G., 'De La Grammatextualité', *Poetique*, 59 (1984), 283–94

Leary, Patrick, 'Googling the Victorians', *Journal of Victorian Culture*, 10.1 (2005), 72–86 <https://doi.org/10.3366/jvc.2005.10.1.72>

Lewis, David D., 'Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval', in *Machine Learning: ECML-98*, ed. by Claire Nédellec and Céline Rouveirol, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer, 1998), pp. 4–15 <https://doi.org/10.1007/BFb0026666>

Liddle, Dallas, 'Reflections on 20,000 Victorian Newspapers: "Distant Reading" The Times Using The Times Digital Archive', *Journal of Victorian Culture*, 17.2 (2012), 230–37 <https://doi.org/10.1080/13555502.2012.683151>

Linstead, Erik, Paul Rigor, Sushil Bajracharya, Cristina Lopes, and Pierre Baldi, 'Mining Concepts from Code with Probabilistic Topic Models', in *Proceedings of the Twenty-Second IEEE/ACM International Conference on Automated Software Engineering - ASE '07* (presented at the the twenty-second IEEE/ACM international conference, Atlanta, Georgia, USA: ACM Press, 2007), p. 461 <https://doi.org/10.1145/1321631.1321709>

Liu, Alan, 'The Meaning of the Digital Humanities', *PMLA*, 128.2 (2013), 409–23 <https://doi.org/10.1632/pmla.2013.128.2.409>

Liu, Dapeng, and Shaochun Xu, 'Challenges of Using LSI for Concept Location', in *Proceedings of the 45th Annual Southeast Regional Conference*, 2007, pp. 449–454

Llewellyn, Clare, Claire Grover, and Jon Oberlander, 'Improving Topic Model Clustering of Newspaper Comments for Summarisation', in *Proceedings of the ACL 2016 Student Research Workshop* (presented at the Proceedings of the ACL 2016 Student Research Workshop, Berlin, Germany: Association for Computational Linguistics, 2016), pp. 43–50 <https://doi.org/10.18653/v1/P16-3007>

———, 'Summarizing Newspaper Comments', in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014 <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8098> [accessed 24 February 2020]

Lo, Ven-Hwei, Anna Paddon, and Hsiaomei Wu, 'Front Pages of Taiwan Daily Newspapers 1952–1996: How Ending Martial Law Influenced Publication Design', *Journalism & Mass Communication Quarterly*, 77.4 (2000), 880–97 <https://doi.org/10.1177/107769900007700410>

Loeb, Lori Anne, *Consuming Angels: Advertising and Victorian Women* (Oxford: Oxford University Press, 1994)

Loftus, Donna, 'Capital and Community: Limited Liability and Attempts to Democratize the Market in Mid-Nineteenth-Century England', *Victorian Studies*, 45.1 (2002), 93–120

Loper, Edward, and Steven Bird, 'NLTK: The Natural Language Toolkit', *ArXiv:Cs/0205028*, 2002 <http://arxiv.org/abs/cs/0205028> [accessed 3 March 2020]

Lord Thomas Babbington Macaulay, 'Minute on Indian Education (1835)', in *Archives of Empire: From The East India Company to the Suez Canal*, by Mia Carter and Barbara Harlow (Durham & London: Duke University Press, 2003), I, 127–38

Mabey, Ben, *PyLDAvis*, version 2.12, 2020 <https://github.com/bmabey/pyLDAvis> [accessed 4 March 2020]

MacKenzie, John M., ed., *Imperialism and Popular Culture* (Manchester University Press, 1986)

———, *Orientalism: History, Theory and the Arts* (Manchester: Manchester University Press, 1995)

———, *Popular Imperialism and the Military: 1850-1950* (Manchester University Press, 1992)

———, *Propaganda and Empire: The Manipulation of the British Public Opinion 1880-1960* (Manchester: Manchester University Press, 1984)

MacLauchlan, Donald J., review of *The Struggle for the Mastery of Europe 1848—1918*, by A. J. P. Taylor, *Weltwirtschaftliches Archiv*, 79 (1957), 66–68

Maguire, Joseph, 'Globalisation, Sport And National Identities: "The Empires Strike Back"?', *Loisir et Société / Society and Leisure*, 16.2 (1993), 293–321 <https://doi.org/10.1080/07053436.1993.10715455>

Mahmud, Athir, Mél Hogan, Andrea Zeffiro, and Libby Hemphill, 'Teaching Students How (Not) to Lie, Manipulate, and Mislead with Information Visualization', in *Big Data Factories*, ed. by S. A. Matei, S. P. Goggins, and N. Jullien (Springer, 2017), pp. 101–114

Malcolm, Dominic, *Globalizing Cricket: Englishness, Empire and Identity*, Globalizing Sports Studies (London and New York: Bloomsbury Academic, 2013)

———, 'Malign or Benign? English National Identities and Cricket', *Sport in Society*, 12.4–5 (2009), 613–28 <https://doi.org/10.1080/17430430802702897>

Malcolm, Dominic, and Philippa Velija, 'Cricket: The Quintessential English Game?', in *Sport and English National Identity in a 'Disunited Kingdom'*, ed. by Tom Gibbons and Dominic Malcolm (London: Routledge, 2017), pp. 18–33

Mangan, James Anthony, *The Games Ethic and Imperialism Aspects of the Diffusion of an Ideal* (London; Portland, Ore.: F. Cass, 1986)

Markovits, Stefanie, *The Crimean War in the British Imagination* (Cambridge: Cambridge University Press, 2013)

Martel, Carol M., 'British Ladies Female Emigrant Society', *Historical Dictionary of the British Empire* (Westport, CT: Greenwood Press, 1996), 189–90

Massie, Robert K., *Dreadnought: Britain, Germany and the Comming of the Great War* (London: Vintage, 2007)

Matzke, Rebecca Berens, *Deterrence Through Strength: British Naval Power and Foreign Policy Under Pax Britannica* (Lincoln and London: University of Nebraska Press, 2011)

McCallum, Andrew K., *MALLET: A Machine Learning for Language Toolkit* (University of Massachusetts Amherst, 2002) <http://mallet.cs.umass.edu>

McCartney, Paul T., *Power and Progress: American National Identity, the War of 1898, and the Rise of American Imperialism* (Baton Rouge: Louisiana State University Press, 2006)

Meeks, Elijah, 'Is Digital Humanities Too Text-Heavy? | Digital Humanities Specialist', *Digital Humanities Specialist*, 2013 <https://dhs.stanford.edu/spatial-humanities/is-digital-humanities-too-text-heavy/> [accessed 5 February 2020]

Megill, Allan, 'Recounting the Past: "Description," Explanation, and Narrative in Historiography', *The American Historical Review*, 94.3 (1989), 627–53 <https://doi.org/10.1086/ahr/94.3.627>

Mervola, Pekka, *Kirja, Kirjavampi, Sanomalehti: Ulkoasukierre Ja Suomalaisten Sanomalehtien Ulkoasu 1771-1994* (Suomen historiallinen seura, 1995), I

Miles, Ellie, 'Characterising the Nation: How T.P. Cooke Embodied the Naval Hero in Nineteenth-Century Nautical Melodrama', *Journal for Maritime Research*, 19.2 (2017), 107–20 <https://doi.org/10.1080/21533369.2017.1405632>

Miller, Amy, 'Clothes Make the Man: Naval Uniform and Masculinity in the Early Nineteenth Century', *Journal for Maritime Research*, 17.2 (2015), 147–54 <https://doi.org/10.1080/21533369.2015.1094984>

Miller, Ian Matthew, 'Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach', *Poetics*, 41.6 (2013), 626–49 <https://doi.org/10.1016/j.poetic.2013.06.005>

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum, 'Optimizing Semantic Coherence in Topic Models', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11 (Edinburgh, United Kingdom: Association for Computational Linguistics, 2011), pp. 262–272

Minard, Charles-Joseph, 'Carte Figurative des Pertes Sucessives en Hommes de l'Armee Francaise dans la Campagne de Russie 1812 - 1813' (Paris: Regnier et Dourdet, 1844), Bibliothèque numérique patrimoniale des ponts et chaussées, Ecole nationale des ponts et chaussées, Fol.10975

Mohr, John W., and Petko Bogdanov, 'Introduction—Topic Models: What They Are and Why They Matter', *Poetics*, 41.6 (2013), 545–69 <https://doi.org/10.1016/j.poetic.2013.10.001>

Morris, Jan, *Farewell to Trumpets: An Imperial Retreat*, Pax Britannica, 2nd edn, 3 vols (London: Faber & Faber, 2012), III

———, *Pax Britannica: The Climax of an Empire*, Pax Britannica, 2nd edn, 3 vols (London: Faber & Faber, 2012), II

Morriss, Roger, *Naval Power and British Culture, 1760–1850: Public Trust and Government Ideology* (London and New York: Routledge, 2017)

Mort, Frank, *Dangerous Sexualities: Medico-Moral Politics in England Since 1830*, 2nd edn (London and New York: Routledge, 2000)

Moura, F., and P. Heitor, 'An Academic Parable: Robert W. Fogel's Raft', 2014

Mudford, Beth, 'Royal Celebrations in the Twenty-First Century: "Cool Britannia" versus "Britannia Ruled the Waves"', in *Identity Discourses and Communities in International Events, Festivals and Spectacles*, ed. by Udo Merkel, Leisure Studies in a Global Era (London: Palgrave Macmillan UK, 2015), pp. 116–34 <https://doi.org/10.1057/9781137394934_6>

Muir, Ramsay, *A Short History of the British Commonwealth*, 4th edn, 2 vols (London & Liverpool: George Philips & Son, 1927)

Mullen, Lincoln, 'Digital Humanities Is a Spectrum, or "We're All Digital Humanists Now"', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 237–38

Murdock, Jaimie, and Colin Allen, 'Visualization Techniques for Topic Model Checking', in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015 <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10007>

Mussell, J., *The Nineteenth-Century Press in the Digital Age* (Springer, 2012)

Mussell, James, *The Nineteenth-Century Press in the Digital Age*, Palgrave Studies in the History of the Media (Basingstoke: Palgrave Macmillan, 2012)

Mussell, James, and Matthew Taunton, 'News Agencies', ed. by Laurel Brake and Marysa Demoor, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 450–51

Mutuvi, Stephen, Antoine Doucet, Moses Odeo, and Adam Jatowt, 'Evaluating the Impact of OCR Errors on Topic Modeling', in *Maturity and Innovation in Digital Libraries*, ed. by Milena Dobreva, Annika Hinze, and Maja Žumer, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2018), pp. 3–14 <https://doi.org/10.1007/978-3-030-04257-8_1>

Mwangi, W., 'The Lion, the Native and the Coffee Plant: Political Imagery and the Ambiguous Art of Currency Design in Colonial Kenya', *Geopolitics*, 7.1 (2002), 31–62

Naeem, U., A.-R. Tawil, and I.I. Kennedy, 'A Dynamic Segmentation Based Activity Discovery through Topic Modelling', in *IET International Conference on Technologies for Active and Assisted Living (TechAAL)* (presented at the IET International Conference on Technologies for Active and Assisted Living

(TechAAL), London, UK: Institution of Engineering and Technology, 2015) <https://doi.org/10.1049/ic.2015.0136>

Naoroji, Dadabhai, *Poverty of India* (London: Vincent Brooks, Day and Son, 1878)

Nelson, Robert K., 'Mining the Dispatch', 2010 <http://dsl.richmond.edu/dispatch/pages/intro> [accessed 26 September 2017]

Newman, David J., and Sharon Block, 'Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper', *Journal of the American Society for Information Science and Technology*, 57.6 (2006), 753–67

Nicholson, Bob, 'Counting Culture; or, How to Read Victorian Newspapers from a Distance', *Journal of Victorian Culture*, 17.2 (2012), 238–46 <https://doi.org/10.1080/13555502.2012.683331>

———, 'Looming Large: America and the Late-Victorian Press, 1862-1902' (unpublished PhD Disseration, University of Manchester, 2012)

———, 'The Digital Turn', *Media History*, 19.1 (2013), 59–73 <https://doi.org/10.1080/13688804.2012.752963>

———, '"You Kick the Bucket; We Do the Rest!": Jokes and the Culture of Reprinting in the Transatlantic Press', *Journal of Victorian Culture*, 17.3 (2012), 273–86 <https://doi.org/10.1080/13555502.2012.702664>

Niklas, Kai, 'Unsupervised Post-Correction of Ocr Errors' (unpublished Master's thesis, Leibniz Universität Hannover, 2010)

Oelke, Daniela, Dimitrios Kokkinakis, and Mats Malm, 'Advanced Visual Analytics Methods for Literature Analysis', in *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '12 (Avignon, France: Association for Computational Linguistics, 2012), pp. 35–44

Orlandi, Tito, 'Reflections on the Development of Digital Humanities' (presented at the DH2019 - Busa Lecture, Utrecht, 2019)

O'Rourke, Kevin Hjortshøf, 'From Empire to Europe: Britain and the World Economy', in *The Cambridge Economic History of Britain*, ed. by Roderick Floud and Paul Johnson, 2 vols (Cambridge: Cambridge University Press, 2014), II, 60–94

Paine, Lincoln, *The Sea and Civilization: A Maritime History of the World* (London: Atlantic Books, 2013)

Palfray, Thomas, David Hebert, Stéphane Nicolas, Pierrick Tranouez, and Thierry Paquet, 'Logical Segmentation for Article Extraction in Digitized Old Newspapers', in *Proceedings of the 2012 ACM Symposium on Document*

Engineering, DocEng '12 (Paris, France: Association for Computing Machinery, 2012), pp. 129–132 <https://doi.org/10.1145/2361354.2361383>

Pastoureau, M., "L'illustration Du Livre: Comprendre Ou Rêver?', in *Histoire de l'édition Française. Tome I. Le Livre Conquérant. Du Moyen Âge Au Milieu Du XVIIe Siècle*, ed. by R. Chartier and H.J. Martin (Paris: Fayard, 1989), pp. 602–28

Patel, Chirag, Atul Patel, and Dharmendra Patel, 'Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study', *International Journal of Computer Applications*, 55.10 (2012)

Piersma, Hinke, and Kees Ribbens, 'Digital Historical Research: Context Concepts and the Need for Reflection', *BMGN-Low Countries Historical Review*, 124.4 (2013), 78–102

Pigeon, Stephan, 'Steal It, Change It, Print It: Transatlantic Scissors-and-Paste Journalism in the Ladies' Treasury, 1857–1895', *Journal of Victorian Culture*, 22.1 (2017), 24–39 <https://doi.org/10.1080/13555502.2016.1249393>

'Political and Social: Notes and Comments', *The Examiner* (London, 2 February 1878), pp. 10-11 (138-139)

Poovey, Mary, *The Financial System in Nineteenth-Century Britain* (Oxford University Press, 2003)

———, 'Writing about Finance in Victorian England: Disclosure and Secrecy in the Culture of Investment', *Victorian Studies*, 45.1 (2002), 17–41

Porter, Andrew, ed., *The Nineteenth Century*, The Oxford History of the British Empire, 5 vols (Oxford: Oxford University Press, 1999), III

Porter, Bernard, *The Absentminded Imperialists: Empire, Society and Culture* (Oxford: Oxford University Press, 2006)

Potapenko, Anna, and Konstantin Vorontsov, 'Robust PLSA Performs Better Than LDA', in *Advances in Information Retrieval*, ed. by Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, and others, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer, 2013), pp. 784–87 <https://doi.org/10.1007/978-3-642-36973-5_84>

Potter, Simon J., 'Empire, Cultures and Identities in Nineteenth- and Twentieth-Century Britain', *History Compass*, 5.1 (2007), 51–71 <https://doi.org/10.1111/j.1478-0542.2006.00377.x>

———, 'Jingoism, Public Opinion, And The New Imperialism', *Media History*, 20.1 (2014), 34–50 <https://doi.org/10.1080/13688804.2013.869067>

———, *News and the British World: The Emergence of an Imperial Press System, 1876-1922* (Clarendon, 2003)

Price, Leah, *How to Do Things with Books in Victorian Britain* (Princeton: Princeton University Press, 2012)

'Prospect', *Computers and the Humanities*, I.1 (1966), 1–2

Protschky, Susie, 'The Colonial Table: Food, Culture and Dutch Identity in Colonial Indonesia', *Australian Journal of Politics & History*, 54.3 (2008), 346–57

Ramos, Juan Enrique, 'Using TF-IDF to Determine Word Relevance in Document Queries', in *Proceedings of the First Instructional Conference on Machine Learning* (presented at the iCML-2003, Piscataway, NJ, 2003) <https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf>

Ramsay, Stephen, 'High Performance Computing for English Majors', 2008 <http://stephenramsay.us/text/2008/04/14/high-performance-computing-for-english-majors/> [accessed 21 October 2016]

———, 'On Building', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 243–45

———, 'Who's In and Who's Out', 2011 <http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/> [accessed 21 October 2016]

———, 'Who's In and Who's Out', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 239–41

Ramsay, Stephen, and Geoffrey Rockwell, 'Developing Things: Notes towards and Epistemology of Building in the Digtital Humanities', in *Debates in the Digital Humanities*, ed. by Matthew K. Gold (Minneapolis: University of Minnesota Press, 2012) <https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities>

'Random Scoring | Elasticsearch: The Definitive Guide [2.x] | Elastic' <https://www.elastic.co/guide/en/elasticsearch/guide/current/random-scoring.html> [accessed 1 February 2019]

Raymond, Joad, 'The Newspaper, Public Opinion, and the Public Sphere in the Seventeenth Century', *Prose Studies*, 21.2 (1998), 109–36 <https://doi.org/10.1080/01440359808586641>

Read, Donald, *The Power of News: The History of Reuters, 1849-1989* (Oxford, New York: Oxford University Press, 1992), p.

Řehůřek, Radim, 'LDA Corpus Topic Composition - Google Groups', *Google Groups*, 2014 <https://groups.google.com/forum/#!searchin/gensim/topic$20percentage$20corpus%7Csort:date/gensim/3cmG23E4Wl4/1zhxiT1d8EIJ> [accessed 2 July 2017]

Řehůřek, Radim, and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, 2010, pp. 45–50

Řehůřek, Radim, and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta: ELRA, 2010), pp. 45–50

Rennie, Jason D. M., Lawrence Shih, Jaime Teevan, and David R. Karger, 'Tackling the Poor Assumptions of Naive Bayes Text Classifiers', in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03 (Washington, DC, USA: AAAI Press, 2003), pp. 616–623

Richardson, John, 'Reader's Letters', in *Pulling Newspapers Apart*, by Bob Franklin (London and New York: Routledge, 2008), pp. 56–66

Robb, George, *White-Collar Crime in Modern England: Financial Fraud and Business Morality, 1845-1929* (Cambridge University Press, 2002)

Rockwell, Geoffrey, 'Inclusion in the Digital Humanities', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 247–53

Röder, Michael, Axel-Cyrille Ngonga Ngomo, Ivan Ermilov, and Andreas Both, 'Detecting Similar Linked Datasets Using Topic Modelling', in *The Semantic Web. Latest Advances and New Domains*, ed. by Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2016), pp. 3–19 <https://doi.org/10.1007/978-3-319-34129-3_1>

Rodger, N.A.M., 'Commissioned Officers' Careers in the Royal Navy, 1690–1815', *Journal for Maritime Research*, 3.1 (2001), 85–129 <https://doi.org/10.1080/21533369.2001.9668314>

Rogowitz, Bernice E., Lloyd A. Treinish, and Steve Bryson, 'How Not to Lie with Visualization', *Computers in Physics*, 10.3 (1996), 268–273

Röhle, Bernhard, and Theo Rieder, 'Digital Methods: Five Challenges', in *Understanding Digital Humanities*, ed. by David M. Berry (London: Palgrave Macmillan UK, 2012), pp. 67–84 <https://doi.org/10.1057/9780230371934_4>

Roland, Lena, and David Bawden, 'The Future of History: Investigating the Preservation of Information in the Digital Age', *Library & Information History*, 28.3 (2012), 220–36 <https://doi.org/10.1179/1758348912Z.00000000017>

Roth, Mitchel P., and James Stuart Olson, *Historical Dictionary of War Journalism* (Greenwood Publishing Group, 1997)

Roy, Tirthankar, *Economic History of India, 1857-1947* (Oxford University Press, 2011)

Rubinstein, William D., 'The World Hegemon: The Long Nineteenth Century, 1832 - 1914', in *A World by Itself: A History of the British Isles*, ed. by Jonathan Clark (London: Pimlico, 2011), pp. 451–565

Rüger, Jan, *Great Naval Game: Britain and Germany in the Age of Empires* (Cambridge: Cambridge University Press, 2007)

———, 'Nation, Empire and Navy: Identity Politics in the United Kingdom 1887-1914', *Past & Present*, 3.185 (2004), 159–87

Rutterford, Janette, David R. Green, Josephine Maltby, and Alastair Owens, 'Who Comprised the Nation of Shareholders? Gender and Investment in Great Britain, c. 1870–1935', *The Economic History Review*, 64.1 (2011), 157–87 <https://doi.org/10.1111/j.1468-0289.2010.00539.x>

Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz, *A Bayesian Approach to Filtering Junk E-Mail*, AAAI Technical Report (Association for the Advancement of Artificial Intelligence, 1998), pp. 55–62 <https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-009.pdf>

Said, Edward W., *Culture and Imperialism* (London: Vintage, 1994)

———, *Orientalism* (London: Penguin, 1991)

Salton, G., A. Wong, and C. S. Yang, 'A Vector Space Model for Automatic Indexing', *Commun. ACM*, 18.11 (1975), 613–620 <https://doi.org/10.1145/361219.361220>

Salton, Gerard, 'Recent Trends in Automatic Information Retrieval', in *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '86 (Palazzo dei Congressi, Pisa, Italy: Association for Computing Machinery, 1986), pp. 1–10 <https://doi.org/10.1145/253168.253171>

Sample, Mark, 'The Digital Humanities Is Not about Building, It's about Sharing', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 255–57

Sanders, Sarah K., 'Assessing the Missions of Digital Humanities Centers', 2019

Schindler, Johanna, and Philipp Müller, 'Design Follows Politics? The Visualization of Political Orientation in Newspaper Page Layout', *Visual Communication*, 17.2 (2018), 141–61 <https://doi.org/10.1177/1470357217746812>

Schofield, Alexandra, M\aans Magnusson, Laure Thompson, and David Mimno, 'Understanding Text Pre-Processing for Latent Dirichlet Allocation', 2017 <https://www.cs.hmc.edu/~xanda/files/winlp2017.pdf> [accessed 10 February 2019]

Scholes, Robert, and Clifford Wulfman, 'Humanities Computing and Digital Humanities', *South Atlantic Review*, 73.4 (2008), 50–66

Sculley, D., and Bradley M. Pasanek, 'Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities', *Literary and Linguistic Computing*, 23.4 (2008), 409–24 <https://doi.org/10.1093/llc/fqn019>

Segel, Edward B., 'A. J. P. Taylor and History', *The Review of Politics*, 26.4 (1964), 531–46

Semmel, Bernard, *The Rise of Free Trade Imperialism: Classical Political Economy the Empire of Free Trade and Imperialism 1750-1850* (Cambridge: Cambridge University Press, 1970)

Shi, Chenggen, and Jie Lu, 'A Text Mining Model by Using Weighting Technology', *AMCIS 2004 Proceedings*, 2004, 228

———, 'Choosing LSI Dimensions by Document Linear Association Analysis', in *Proceedings of the International Conference on Information and Knowledge Engineering*, 2003

Shirley, Michael, ''Renold's Newspaper', ed. by Marysa Demoor and Laurel Brake, *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (London: British Library, 2009), pp. 539–41

Sidaway, James D, 'The Dissemination of Banal Geopolitics: Webs of Extremism and Insecurity', *Antipode*, 40.1 (2008), 2–8 <https://doi.org/10.1111/j.1467-8330.2008.00568.x>

Silberstein-Loeb, Jonathan, 'The Political Economy of Media', in *The Routledge Companion to British Media History*, ed. by Martin Conboy and John Steel (Routledge, 2014), pp. 75–84

Sinclair, Stéfan, Stan Ruecker, and Milena Radzikowska, 'Information Visualization for Humanities Scholars', 2013 <https://doi.org/10.1632/lsda.2013.6>

Sissons, Ric, and Brian Stoddart, *Cricket and Empire: The 1932-33 Bodyline Tour of Australia*, Routledge Library Editions: Sports Studies, 2nd edn (London: Routledge, 2014)

Slauter, Will, 'Introduction: Copying and Copyright, Publishing Practice and the Law', *Victorian Periodicals Review*, 51.4 (2018), 583–96

Smith, David A., Ryan Cordell, and Elizabeth Maddock Dillon, 'Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers', in *2013 IEEE International Conference on Big Data*, 2013, pp. 86–94 <https://doi.org/10.1109/BigData.2013.6691675>

Snider, Paul B., '"Mr. Gates" Revisited: A 1966 Version of the 1949 Case Study', *Journalism Quarterly*, 44.3 (1967), 419–27 <https://doi.org/10.1177/107769906704400301>

Snow, C. P., *The Two Cultures*, 4th edn (Cambridge: Cambridge University Press, 2012)

Solberg, Janine, 'Googling the Archive: Digital Tools and the Practice of History', *Advances in the History of Rhetoric*, 15.1 (2012), 53–76 <https://doi.org/10.1080/15362426.2012.657052>

Springhall, John O., '"Up Guards and At Them!": British Imperialism and Popular Art, 1880-1914', in *Imperialism and Popular Culture*, by John M. MacKenzie (Manchester: Manchester University Press, 1986), pp. 49–72

Stalmaszczyk, Piotr, ed., *Philosophy of Language and Linguistics: The Legacy of Frege, Russell, and Wittgenstein*, Philosophische Analyse (Boston, [Massachusetts]: De Gruyter, 2014), LIII

Standage, Tom, *The Victorian Internet*, 3rd edn (London and New York: Bloomsbury, 2014)

Steffen, Charles G., 'Newspapers for Free: The Economies of Newspaper Circulation in the Early Republic', *Journal of the Early Republic*, 23.3 (2003), 381–419 <https://doi.org/10.2307/3595045>

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski, and David Buttler, 'Exploring Topic Coherence over Many Models and Many Topics', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12 (Jeju Island, Korea: Association for Computational Linguistics, 2012), pp. 952–961

Steyvers, Mark, and Tom Griffiths, 'Probabilistic Topic Models', *Handbook of Latent Semantic Analysis*, 427.7 (2007), 424–440

Summerfield, Penny, 'Patriotism and Empire: Music-Hall Entertainment, 1870-1914', in *Imperialism and Popular Culture*, by John M. MacKenzie (Manchester: Manchester University Press, 1986), pp. 17–48

Sumpter, Randall S., '"Practical Reporting": Late Nineteenth-Century Journalistic Standards and Rule Breaking', *American Journalism*, 30.1 (2013), 44–64 <https://doi.org/10.1080/08821127.2013.767686>

Sutch, Richard, 'The Treatment Recieved by American Slaves: A Critical Review of the Evidence Presented in Time on the Cross', *Exploraions in Economic History*, 12.4 (1975), 335–438

Tabata, Tomoji, 'Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction', in *Proceedings of the 7th Conference of Japanese Association for Digital Humanities* (presented at the JADH 2017, Doshisha, 2017), pp. 73–78

Tahmasebi, Nina, 'A Study on Word2Vec on a Historical Swedish Newspaper Corpus', in *DHN*, 2018

Tanner, Simon, Trevor Muñoz, and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', *D-Lib Magazine*, 15.7/8 (2009) <https://doi.org/10.1045/july2009-munoz>

Taylor, A.J.P, *English History 1914-1945*, The Oxford History of England, XV, 3rd edn (Oxford: Oxford University Press, 1976)

———, *Germany's First Bid for Colonies, 1884-1885: A Move in Bismarck's European Policy* (Basingstoke: Macmillan, 1938)

Taylor, Antony, '"Some Little or Contemptible War upon Her Hands": Reynolds's Newspaper and the Empire', in *G.W.M. Reynolds: Nineteenth-Century Fiction, Politics, and the Press*, ed. by Anne Humphrey and Lois James (London: Routledge, 2017), pp. 98–120

Temple, Mick, *The British Press* (Maidenhead: McGraw-Hill Education (UK), 2008)

Terras, Melissa, Julianne Nyhan, Edward Vanhoutte, Julianne Nyhan, and Edward Vanhoutte, *Defining Digital Humanities : A Reader* (Routledge, 2016) <https://doi.org/10.4324/9781315576251>

Therón Sánchez, Roberto, Alejandro Benito Santos, Rodrigo Santamaría Vicente, and Antonio Losada Gómez, 'Towards an Uncertainty-Aware Visualization in the Digital Humanities', *Informatics*, 6.3 (2019), 31 <https://doi.org/10.3390/informatics6030031>

Therón Sánchez, Roberto, Antonio Losada Gómez, Alejandro Benito Santos, and Rodrigo Santamaría Vicente, 'Toward Supporting Decision-Making under Uncertainty in Digital Humanities with Progressive Visualization', in *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18 (Salamanca, Spain: Association for Computing Machinery, 2018), pp. 826–832 <https://doi.org/10.1145/3284179.3284323>

Thompson, Edward P., *The Making of the English Working Class* (London: Penguin, 1963)

Thompson, J. Lee, 'Fleet Street Colossus: The Rise and Fall of Northcliffe, 1896-1922', *Parliamentary History*, 25.1 (2006), 115–38 <https://doi.org/10.1353/pah.2006.0011>

Torget, Andrew J., Rada Mihalcea, Jon Christensen, and Geoff McGhee, 'Mapping Texts: Combining Text-Mining and Geo-Visualization to Unlock the Research Potential of Historical Newspapers', *University of North Texas Digital Library*, 2011 <https://pdfs.semanticscholar.org/4b40/d6b77b332214eefc7d1e79e15fbc2d86d

Tosh, John, *The Pursuit of History: Aims, Methods and New Directions in the Study of History*, 6th edn (London anNew York: Routledge, 2015)

Tosh, Nick, 'Science, Truth and History, Part I. Historiography, Relativism and the Sociology of Scientific Knowledge', *Studies in History and Philosophy of Science Part A*, 37.4 (2006), 675–701 <https://doi.org/10.1016/j.shpsa.2006.09.004>

Trudeau, Noah A., 'A Naval Tragedy's Chain of Errors', *Naval History Magazine*, 24.1 (2010) <https://www.usni.org/magazines/naval-history-magazine/2010/february/naval-tragedys-chain-errors>

Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, 'Robustness May Be at Odds with Accuracy', *ArXiv:1805.12152 [Cs, Stat]*, 2019 <http://arxiv.org/abs/1805.12152> [accessed 25 February 2020]

Tucker, Albert V., 'Army and Society in England 1870-1900: A Reassessment of the Cardwell Reforms', *Journal of British Studies*, 2.2 (1963), 110–41

Turner, Michael, 'Enclosures in Britain 1750–1830', in *The Industrial Revolution A Compendium*, ed. by L. A. Clarkson, Studies in Economic and Social History (London: Macmillan Education UK, 1990), pp. 211–95 <https://doi.org/10.1007/978-1-349-10936-4_4>

Tworek, Heidi J. S., 'Political and Economic News in the Age of Multinationals', *Business History Review*, 89.3 (2015), 447–74 <https://doi.org/10.1017/S0007680515000677>

Underwood, Ted, 'Theorizing Research Practices We Forgot to Theorize Twenty Years Ago', *Representations*, 127.1 (2014), 64–72 <https://doi.org/10.1525/rep.2014.127.1.64>

Unsworth, John, 'The State of the Digital Humanities', 2010 <http://www.people.virginia.edu/~jmu2m/state.of.dh.DHSI.pdf>

———, 'What Is Humanities Computing and What Is Not?', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 35–48

Van Loan, Charles F., 'Generalizing the Singular Value Decomposition', *SIAM Journal on Numerical Analysis*, 13.1 (1976), 76–83 <https://doi.org/10.1137/0713009>

Vanhoutte, Edward, 'The Gates of Hell: History and the Definition of Digital | Humanities | Computing', in *Defining Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (Farnham: Ashgate, 2013), pp. 119–56

Verbeek, Georgi, 'De wording van een Natie: Duitsland tijdens de lange 19e eeuw', in *Een Geschiedenis van Duisland: Sporen en Dwaalsporen van een Natie* (Leuven and The Hague: Acco, 2010), pp. 105–53

Walker, Daniel D., William B. Lund, and Eric K. Ringger, 'Evaluating Models of Latent Document Semantics in the Presence of OCR Errors', in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10 (Cambridge, Massachusetts: Association for Computational Linguistics, 2010), pp. 240–250

Walker, Daniel, Eric Ringger, and Kevin Seppi, 'Evaluating Supervised Topic Models in the Presence of OCR Errors', ed. by Richard Zanibbi and Bertrand Coüasnon (presented at the IS&T/SPIE Electronic Imaging, Burlingame, California, USA, 2013), p. 865812 <https://doi.org/10.1117/12.2008345>

Walton, J.K., *The English Seaside Resort: A Social History, 1750-1914* (Leicester: Leicester University Press, 1983)

Ward, Paul, *Britishness Since 1870* (London and New York: Routledge, 2004)

Weber, Brooke, '"A Mad Proceeding": Mid-Nineteenth-Century Female Emigration to Australia' (unpublished PhD Dissertation, Royal Holloway, University of London, 2018)

<https://pure.royalholloway.ac.uk/portal/files/31011848/2018weberbphd.pdf.pdf>

Wells, John, *The Royal Navy: An Illustrated Social History, 1870-1982*, New edition edition (Stroud: Sutton Publishing Ltd, 1996)

White, David Manning, 'The "Gate Keeper": A Case Study in the Selection of News', *Journalism Quarterly*, 27.4 (1950), 383–90 <https://doi.org/10.1177/107769905002700403>

Wickham, Hadley, 'Tidy Data', *Journal of Statistical Software*, 59.10 (2014), 1–23

Wiener, Joel H., 'The Nineteenth Century and the Emergence of a Mass Circulation Press', in *The Routledge Companion to British Media History*, ed. by Martin Conboy and John Steel (Routledge, 2014), pp. 206–14

Wilkinson, Leland, and Michael Friendly, 'The History of the Cluster Heat Map', *The American Statistician*, 63.2 (2009), 179–84 <https://doi.org/10.1198/tas.2009.0033>

Willems, Gertjan, 'Digitale Tools Voor Kwalitatieve Data-Analyse Binnen Historisch Comminicatiewetenschappelijk Onderzoek: Toepassingen En Reflecties', *Tijdschrift Voor Communicatiewetenschap*, 45.3 (2017), 170–83

Williams, Francis, *Dangerous Estate: The Anatomy of Newspapers* (Longmans, Green, 1957)

Williams, Kevin, *Read All about It! A History of the British Newspaper* (London and New York: Routledge, 2010)

Williams, Walter L., 'United States Indian Policy and the Debate over Philippine Annexation: Implications for the Origins of American Imperialism', *The Journal of American History*, 66.4 (1980), 810–31 <https://doi.org/10.2307/1887638>

Williamson, Jeffrey G, 'The Impact of the Corn Laws Just Prior to Repeal', *Explorations in Economic History*, 27.2 (1990), 123–56 <https://doi.org/10.1016/0014-4983(90)90007-L>

Williamson, Philip, *National Crisis and National Government: British Politics, the Economy and Empire, 1926-1932* (Cambridge University Press, 2003)

Wills, Craig E., and Doruk C. Uzunoglu, 'What Ad Blockers Are (and Are Not) Doing', in *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, 2016, pp. 72–77 <https://doi.org/10.1109/HotWeb.2016.21>

Wilson, Jobin, Santanu Chaudhury, and Brejesh Lall, 'Improving Collaborative Filtering Based Recommenders Using Topic Modelling', in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and*

Intelligent Agent Technologies (IAT), 2014, I, 340–46 <https://doi.org/10.1109/WI-IAT.2014.54>

Wilson, Kathleen, ed., *A New Imperial History: Culture, Identity and Modernity in Britain and the Empire 1660-1840* (Cambridge: Cambridge University Press, 2004)

Winters, Jane, and Steve F. Anderson, 'Digital History', in *Debating New Approaches to History*, ed. by Marek Tamm and Peter Burke (London: Bloomsbury, 2018), pp. 277–300

van Wissen, Leon, 'Topic Modelling "De Gids": An Explorative Study into the Use of Topic Modelling on a Cultural Periodical' (unpublished Research Master Thesis, Vrije Universiteit, 2019) <https://www.leonvanwissen.nl/publication/vanwissen-2019-degids/vanwissen-2019-degids.pdf>

Wittgenstein, Ludwig, *Philosophical investigations*, 2nd ed. (Oxford: Blackwell, 1958) <http://capitadiscovery.co.uk/edgehill/items/14404> [accessed 23 January 2019]

Woollacott, Janet, Michael Gurevitch, Open University, and James Curran, *Mass Communication and Society* (London: Edward Arnold, 1977) <http://capitadiscovery.co.uk/edgehill/items/69710> [accessed 2 February 2018]

Wrigley, C. J., *A.J.P. Taylor: radical historian of Europe* (London: I.B. Tauris, 2006)

Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea, 'Topic Modeling on Historical Newspapers', in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Stroudsburg, PA, USA: Association for Computational Linguistics, 2011), pp. 96–104

Zahedieh, Nuala, 'Overseas Trade and Empire', in *The Cambridge Economic History of Britain*, ed. by Roderick Floud and Paul Johnson, 2 vols (Cambridge: Cambridge University Press, 2014), I, 392–420

Appendix 1: Examples of OCR Quality

article ID: WO1_LEMR_1872_08_08-0008-028

'Crips antb Wottrs. N ORTH-EABTrE R 1 RAILWAY. YOnKS11IRE AGI1UR1LTUC L. I{JIW AT UNLID.'N. On Frldalr. Augl'" MIh, a1 speial ttairn 1W l . .j . etl rt s !. n m.. for NALTO.N, cr1.ig1 I .-, Fltr3 ii otwrc fr.iI. iMt, 1;3 ut I0 p m. 1y ftR)EHL Yoik, Antsut, tbF,. _____„ NORTH-EA8TER N BAIL W &I A1UtANGEWLNT' f7r th-. 1MtE of FIR'fT, S'OS\D, .ni TiMliRD CLA.AbS VOURlT Tl:ICP:F' F il t bein torm. fr'om Me4; l3t', *.).t tht.At"r. 172. ;ir LIU~k1. 6 ISX id.N~ <in-z; !1,ly Tork, why, lIt. 4.43 rLVp1c SH EtL, YO a 1bk H Ii.AIItI.I lc Y A ltemorn, Iotn Lo-ts. Ha.3i. Iallf... I kI.,.n. Nennr, Ottlesaof orc-touL 11.2 I," :. ~t').n V~x"ct wiaFtbri tr-h: &bo r,d vi -trz tFT ,ai Uir.4-r ?lUblONS TO B.,A(RFOtVL. S{)Tfif3Oltl, :iUN FLLL2-. WOOD fo L ,o,,rm 11011 iBoe~dafl iontfrtg trtlns lire cow r~n d..lrtlfr1nz t; .V. Jlndl.- Hidt t uId, mmhJtsitd 5o"o.br, DitJd1 .-hr--'j *, ., CIAd<nefSHIRE B v~ltNl IiL4 iisnc,,j, SL. .. It 'i Ls e-s, tport, antd i1~Lw~~5O'd r JiMno 13th. oS71au4 tCa * " i *tnlayo r to .3,J SP to 1ck..l, AL.~I3.i~i i.. lb.E *. 1 3L~ c'Ir .u : Ppdraltmtrn s : ny wititr-das a rr.on ct, t- V": ..lt* 3 a.la, frorn id.otisort .r t J.533N p-x.tt L ;,- .. ,; . A~dtloCeI I Xto.; T,0.I'. O~ld S1 l33 3st L; * , i'; Q. 1- rootbsortatG.lo iLm. ak iJD;3)h, ; 1,rst and aoewn-4oLIa phmongorsa bl, ' Xk(- tb . -i, tmt-mr *1 o-dinary fara. di' OtLr.1t. Cla of Srnperlttent 0 thero r .' ia itt a 8 Uvwdibeater. Jutu 7Ttb. 1"7!. Il 11ii7 LAINCA13a IRE A20.) YU01.ki6 i U fl.1L; bAFPzrNOW EXUhtliSYoNS TO LrrT'L ilJU3H Ii.) !1 HEOLsJ.rNvslTif 1AKF On Saturday, .7r ,c 11?,t, ib,, %ntl .i r .rl.32 " I.a3 ft" MUt. ao~e n bsc21J 'i tain 7rill 1,1 rwn t 1.1. 1 Lb S.112 fiJU'ti I ~ot s HI..l1Iot23 lsb. iLado, tlc--.: JJIlI.i., *irn14, nir..1'. v ~A1. -:- HUWititSt.,l~~i~~i~ii~$',~:::-t.]t-r : Z-u,7., aSkd unlrimnlitI r ttiorLt 131r and all i'L ormnstlg. can be obtnj.l1 at the S- ..Uo.:A OtLbe Of at V Sntarlntmidct of Me Lne. Vicr3,zL r tria t.u ftLe,l1z Ji.:a 16ti. 1Is72. 'I inn 1 8LEL (I M ItAN.-RtY &L MAIL - Vl'g..x:,IdR tc LOUGLAS aitly tiYnud3s. .:crs.Il Iron. ! 4n f 'lea beamC, aitr1l. at ure Ur4:--n..,,I r.- .uu fr...o I ax) h. e 1!.du13 g a t.11 alA S]-.JI.:D!. In J.-; aIA .La ..a steamer will l-,av, lIon;me for 1_4C1-oelt 1.n3rwarD Laverpcol to Z1iayna I!frevl, ec-ry Xonmtay at mo.. ,L 1~ Io V m wh Priduy. TcLr1emUx;LR "-1.7> djt.>:l btlit r I)Dsgai. let June, lb.2. J-iO.J.GO Diill;iU, -'-- f'LA86iOW and the H11HLAN'I)J.-t;.,..t . G m (.LInu i.d call 0011 OAni. y . . ... .1... . hrowGl~gow ht7 h m, and Gr,.:-a-k C t Snen, Sor 0U., ruo.t W7ishn, azd 1In-mlll .d d..ily GleamoN Gairll b. P~a-~ir ."!l:^ Iona. MaEia > -': v ti Weit lIlasilend5, Pee billL„ 1.131. i, p wi wu.-.i %-- I. U; *.. vd D0ft . BJok!4lPr, 1, MN::Xg~-te- Lr. !1. -...--_ atiorer h.r !d Ilrmt.3er, 64, ;ortt. ,:rr ,AiI: or 1b P -2 tWun Wo Datrd UutrL esori ard C.. 1.', Igjt-o'italty IGL..-. 1 B OLT{JN ABBEY, th11 ST IDi iu1.t -% . dO ke. To Pl anre and Piceicl 'aT`.' 9 .!' Btation, en CD plaioalf tO JOUN lJOUtBoi, Jun Cr.; V. '* ili,. .5 1 &t T R U N Ki , BATIM 1UNNI-£ BV A.;.' T ('.ito awl. AP.1 iT t)tCK II..1 3.fl* ir llKT 1, a ndi l 1. P.nOOl .1 ILnWE-ELALNE. 1.'~ E r .!i 2 .:

XTISITURS to th- biA-S! DE Artd T .L. 0 .rUre . I. TiiA; -.i l;N . ; : v A:t .L. il.Nlek1:t'2f. Menor.:or: r r-;.oir:* >.N.. o1 s'ji :- Il 1 .1:

article ID: WO1_BNWL_1870_01_06-0003-014

CAPTURING FISH BY MOONLIGHT.

OVERTAKEN by night, when travelling through the SI Jura Mountains in France, many years ago, we stopped for supper and a night's lodging at a small Ii wayside inn of rather dilapidated appearance, under w some apprehension as to how we should fare there; P but the cordial greeting of the landlord, whose fat ti rubicund face and moist twinkling blue eye, gave co promise of good entertainment, soon dispelled our t(fears. One of his first questions was whether we liked trout for supper fresh from the brook. Of to course we did; but, to our astonishment-it being ' now pitch dark-we learned that the fish had yet t to be caught. Being an ardent angler, and curious to learn how the thing was to be done, upon his invitation, we accompanied him to the scene of action, h a brawling mountain brook, within a few yards of n hisdoor. Before starting he took from a closet, where s it had been stowed away, an ordinary glass globe c lantern, with two long tin tubes fixed to it on either side, through which the flxme was supplied h with air. Lighting it, he thcn took from his pocket 0 a common pruning knife, with a hawk's bill, called a " serp2;" he was now prepared for the fray. A t walk of two or three minutes brought us to the side : of a deep dark pool, which, with the glare of the P lamp dancing upon it like a " will outhe-wisp," c lo ked like the bottomless pit. With the queer- looking lantern in his leit hand and theright armed a with the formidable knife, the landlord seated him- u self on as flat, projecting rock whence the desetnt r was sheer to tue bottom of the pool. He then slowly (thrust the still burning lamp into the deep water, s where it looked like a great, glowing kohinoor. a Holding it thus for- about a minute, he raised it t evenly and slowly to the surface, and around it, to our delight aid astonishment, were fifteen or twenty fish of different sizes pressing their noses against the glass as if eager to get at the light. Then, selecting the best fish, our host adroitly tapped four of them on the head witb the bill of his knife. They tuined on their sides dead without even a flutter. Thus in leFs time than it takes to write this account we had four prime, half-pound trout, which, with the addition of an omelette au lard, such as the French only can make, a mound of perfumed golden mountain i butter, and a bottle of Baune, covered with the dust of quarter of a century, we had a supper worthy of record in Brillar-Savarin's immortal Pl',siofogie di Gout. This pleseant little adventure at the wayside inn was recalled to cur memory by a statement in one of our exchanges. It seems that a light is quite as attractive to the fish of the great deep as to their cousins of the mountain brooks. The professional fishermen on the coast of France, having recently discovered this fact, are now making heavy draughts of fish, attracted to their nets by powerful submerged lights.- Turt;, FPied, and Farm.